This is a digital copy of a book that was preserved for generations on library shelves before it was carefully scanned by Google as part of a project to make the world's books discoverable online.

It has survived long enough for the copyright to expire and the book to enter the public domain. A public domain book is one that was never subject to copyright or whose legal copyright term has expired. Whether a book is in the public domain may vary country to country. Public domain books are our gateways to the past, representing a wealth of history, culture and knowledge that's often difficult to discover.

Marks, notations and other marginalia present in the original volume will appear in this file - a reminder of this book's long journey from the publisher to a library and finally to you.

Usage guidelines

Google is proud to partner with libraries to digitize public domain materials and make them widely accessible. Public domain books belong to the public and we are merely their custodians. Nevertheless, this work is expensive, so in order to keep providing this resource, we have taken steps to prevent abuse by commercial parties, including placing technical restrictions on automated querying.

We also ask that you:

+ *Make non-commercial use of the files* We designed Google Book Search for use by individuals, and we request that you use these files for personal, non-commercial purposes.

+ *Refrain from automated querying* Do not send automated queries of any sort to Google's system: If you are conducting research on machine translation, optical character recognition or other areas where access to a large amount of text is helpful, please contact us. We encourage the use of public domain materials for these purposes and may be able to help.

+ *Maintain attribution* The Google "watermark" you see on each file is essential for informing people about this project and helping them find additional materials through Google Book Search. Please do not remove it.

+ *Keep it legal* Whatever your use, remember that you are responsible for ensuring that what you are doing is legal. Do not assume that just because we believe a book is in the public domain for users in the United States, that the work is also in the public domain for users in other countries. Whether a book is still in copyright varies from country to country, and we can't offer guidance on whether any specific use of any specific book is allowed. Please do not assume that a book's appearance in Google Book Search means it can be used in any manner anywhere in the world. Copyright infringement liability can be quite severe.

About Google Book Search

Google's mission is to organize the world's information and to make it universally accessible and useful. Google Book Search helps readers discover the world's books while helping authors and publishers reach new audiences. You can search through the full text of this book on the web at `http://books.google.com/`

Math 1009.06.3

HARVARD COLLEGE
LIBRARY

FROM THE BEQUEST OF

JAMES WALKER
(Class of 1814)

President of Harvard College

"Preference being given to works in the Intellectual and Moral Sciences"

SCIENCE CENTER LIBRARY

Digitized by Google

Digitized by Google

ARCHIVES OF PHILOSOPHY
PSYCHOLOGY AND SCIENTIFIC METHODS

Editorial communications should be addressed to Professor J. McKeen Cattell, Garrison, N. Y., or to Professor Frederick J. E. Woodbridge, Columbia University, New York City.

Subscriptions and advertisements should be sent to The Science Press, Sub-Station 84, New York City. The subscription price is five dollars a volume containing between six and seven hundred pages. The numbers are as follows:

1. Measurements of Twins: Edward L. Thorndike. 50 cents.
2. Avenarius and the Standpoint of Pure Experience: Wendell T. Bush. 75 cents.
3. The Psychology of Association: Felix Arnold. 50 cents.
4. The Psychology of Reading: Walter F. Dearborn. $1.00.
5. The Measurement of Variable Quantities: Franz Boas. 50 cents.
6. Linguistic Lapses: F. L. Wells. $1.00.
7. The Diurnal Course of Efficiency: H. D. Marsh. 90 cents.
8. The Time of Perception as a Measure of Differences in Sensations; Vivian A. C. Henmon. 60 cents.
9. The Psychology of Mentally Deficient Children: Naomi Norsworthy. *In press.*

THE JOURNAL OF PHILOSOPHY
PSYCHOLOGY AND SCIENTIFIC METHODS

The contents of recent numbers include:

The Knowledge Experience and its Relationships: John Dewey.
Kant's Doctrine of the Basis of Mathematics: Josiah Royce.
The Psychical Complex called an Interest: Lucinda Pearl Boggs.
The Issue between Idealism and Immediate Empiricism: Charles M. Bakewell.
Mental Elements of Dreams: Will S. Monroe.
Feeling and Conception: Kate Gordon.
Cognitive Experience and its Object: B. H. Bode.
Association and Atomism: Felix Arnold.
A Philosophical Confession: Harald Höffding.
A Syntactician among the Psychologists: Basil L. Gildersleeve.
The Real and the Pseudo Psychology of Religion: Irving King.
The Reeducation of an Aphasic: Shepherd Ivory Franz.
An Apparent Contradiction in the Modern Theory of Judgment: W. B. Pillsbury.
An Empirical Definition of Consciousness: Wendell T. Bush.
How Two Minds can know One Thing: William James.
Some Outstanding Problems for Philosophy: Cassius J. Keyser.
Immediate Empiricism: John Dewey.

$3 a Year (26 numbers) **15 cents a Copy.**

LIBRARY OF PHILOSOPHY
PSYCHOLOGY AND SCIENTIFIC METHODS

Theory of Mental and Social Measurements: Edward L. Thorndike. $1.50.
Science and Hypothesis: Henri Poincaré. Translated by George Bruce Halsted, with an Introduction by Josiah Royce. $1.50.

THE SCIENCE PRESS,
Sub-Station 84, New York City.

Digitized by Google

Math 1009.06.3

THE MEASUREMENT OF VARIABLE QUANTITIES

BY

FRANZ BOAS, PH.D.

Professor of Anthropology, Columbia University

ARCHIVES OF

PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS

EDITED BY

J. McKEEN CATTELL AND FREDERICK J. E. WOODBRIDGE

No. 5, JUNE, 1906

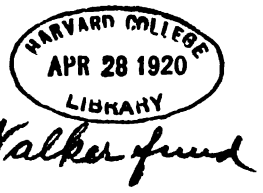
Columbia University Contributions to Philosophy and Psychology, Vol. XIV. No. 2

NEW YORK

THE SCIENCE PRESS

Digitized by Google

Math 1009.06.3



Walker fund

PRESS OF
THE NEW ERA PRINTING COMPANY
LANCASTER, PA.

Digitized by Google

PREFACE

THE present treatise contains the introduction to a course on the statistical treatment of biological and psychological measurements, which I have given for ten years at Columbia University. The form selected for the demonstration of the principles of measurement of variables was chosen on account of the limited mathematical preparation of students who have devoted themselves to the study of anthropology, biology, and psychology, which made it necessary to avoid, so far as feasible, all application of the calculus.

While the book was in the hands of the printer, the "Wahrscheinlichkeitsrechnung und Kollektivmasslehre," by Heinrich Bruns, was published. It to a great extent follows methods similar to those used by me, and is much more comprehensive. Nevertheless I have decided to publish my treatment of the subject, because it seems adapted to the peculiar needs of American students.

FRANZ BOAS.

NEW YORK, May, 1906.

Digitized by Google

CONTENTS

Digitized by Google

The Measurement of Variable Quantities

I. INTRODUCTORY; CONSTANTS AND VARIABLES

§ 1

In the quantitative study of nature we distinguish two separate classes of objects and phenomena — constants and variables. The former give the same quantitative results whenever they are measured under exactly the same conditions; the latter give different quantitative results when measured at different times, because the governing conditions are complex and never quite the same.

When we measure the length of a metal rod, we consider it as the same identical object whenever measured, and therefore we assume that it must always have the same length, provided the variable conditions affecting its length remain the same. Only when changes occur that modify permanently its inner structure, and by which its identity is changed, do we speak of a change in the length of the rod. If our refined measurements give different results for the length of the rod, we ascribe these to lack of control of conditions, and we call them 'errors of observation.'[1] If, on the other hand, we cast a number of rods in the same mold and as nearly as possible under the same conditions, the slight differences of conditions in casting will become permanent characteristics of the several rods, and the errors of manufacture of each individual specimen will make them a series, intended to represent the same kind of an object, but, owing to individual differences, variable.

In the same way, if we wish to determine the weight of a cubic centimeter of pure iron, the weight appears as a constant, because both a cubic centimeter and pure iron are identically the same in all cases. Each individual experiment will therefore be an approach to this constant weight, affected by errors of observation according to changing physical condition, to inaccuracies in the size of the cubic centimeter, and to impurities in the iron. On the other hand, the same

[1] The term 'error of observation' is applied to constants, and is generally used in a restricted sense, signifying the error due to the technique of measurement and determining the limits within which differences, if they should occur, can no longer be recognized.

Digitized by Google

series of iron cubes are, when considered individually and objectively, the variable representatives of what is intended to be a cubic centimeter of pure iron.

It appears from these examples, that — according to the principle of identity — we consider a phenomenon which is completely defined as always the same, and therefore as a constant. As soon as any of the constituent or controlling elements are no longer completely defined, the phenomena are no longer identical, but separate individualities. Differences of the results of measurements are called, in the first group, 'differences due to errors of observation'; in the second group, 'variates'; and the phenomena are called, respectively, 'constants' and 'variables.'

While in some groups of phenomena a complete definition can be given which compels us to consider a repetition of a phenomenon as identical with the original one, in others no such definitions are possible, and the individual repetitions always possess independent elements which are not contained in their common definition. These phenomena must always be considered as variables, and their common definition is that of a class embracing the individual phenomena.

The identity of two objects or phenomena is inferred partly from considerations that have no relation to measurements, and we conclude that, on account of their identity, the measurements of the two objects or phenomena must be the same; but we also conclude conversely that when the measurements are not the same the objects or phenomena cannot be identical, and also that when they have the same measurements — that is, when they are constant — they are identical. This last conclusion is, of course, empirical and open to refutation by new facts. For this reason it may be that with increasing knowledge objects or phenomena which once appeared as constants may come to be considered as variables, because what seemed at one time as quantitatively the same is proved to be different; or what seemed at one time identical is proved to contain different elements. In other words, certain parts of the error of observation may be proved to be due to individual differences between observations. The discovery of variations in latitude and of new elements which are found in very small quantities mixed with other elements illustrates this point.

Strictly speaking, no two measurements are absolutely the same. If, in our definition of the phenomenon measured, the differences between repetitions are taken into account, it will be a variable; if they are disregarded, and only the elements common to all repetitions

Digitized by Google

are included, it will be a constant. The example given before illustrates that a cubic centimeter of pure iron is a constant, but that the individual cubes measured may also be considered as variates. A cubic centimeter of pure iron, under given physical conditions, is completely defined, and therefore identically the same whenever measured. The individual cube does not quite correspond to our definition, and, in so far as its individual peculiarities are considered as vitiating our definition, they are errors of observation. When these individual peculiarities are considered as part of the definition, the same cubes would become representatives of the series of cubic centimeters of iron as they exist, and in this sense the individual peculiarities would be considered as variates.

In empirical studies, identity or diversity is often inferred from sameness or diversity of measured values. We have seen that absolute sameness of measurement does not exist. It is therefore a matter of judgment whether a quantity shall be considered as constant or as variable. While considerable differences will always lead to the conclusion that our definition is incomplete — in other words, that the quantity measured is variable — no fixed lowest limit can be given under which variations must be disregarded, as long as they are discernible by means of the standard used in making the measurement. Our decision will always depend upon the question whether we consider each quantity individually, or as adequately represented by what is common to all the quantities measured. Thus, when we compare a number of fairly uniform centimeter rods, and by abstraction define the length of the centimeter as the type of our rods, the measurable deviations will be considered as errors of observation. When, on the other hand, the differences are found to be so great that the centimeter can no longer be considered as the type of the series, each rod must be considered as an individual, and the rods present a series of variates.

It will be noticed that in comparisons between two series quantitative variations in each series will be the more negligible, the greater the differences between the series. Thus, in comparing a series of inaccurate lengths of centimeters and meters, we are more ready to recognize the two types and to consider their deviations as errors, than in a case where we compare inaccurate lengths of one centimeter and of two centimeters.

It appears from what has been said that the essential difference between constants and variables consists in the fact that the constant is an individual considered as a complete representation of a class,

Digitized by Google

the definition of the class and of the individual being the same, and that a variable is a series of individuals of the same class, but each individual considered as different from the other. In the former case, differences of individual measurements are considered as due to subordinate causes; in the second case, these subordinate causes and the causes determining the type are given equal importance.

§ 2

Since a variable consists of a series of individuals constituting a class, the measurement of a variable, to be complete, must consist of a series of measurements of all the individuals of the class. Two variables will appear to us as the same when the series of measurements representing each is the same. In this case we may infer that both variables represent the same class and the same groups of imperfectly known modifying conditions. When we find two variables that fulfil these conditions, we consider them as identical; that is to say, we form the abstraction of the existence of a class of phenomena, the individual cases of which are distributed according to a certain law, and each series of measurements is considered as following this law. Thus the measurement of the variable may be reduced to the determination of those elements which determine the general law of distribution. When these elements are determined from a single limited series of observations, they may be expected to differ from those representing the abstract type of distribution which would be obtained, could the whole infinitely large number of individuals be observed. Since the single series are considered as identical with the type, *i. e.*, the complete infinitely long series of individuals, we call the differences between the individual values and the ideal values 'errors' in the same way as we call 'errors' the differences between one centimeter and individual rods, intended to be one centimeter in length, but differing from this value on account of imperfections of manufacture.

While in nature the number of cases constituting a variable series is limited, we conceive the series in forming the abstract law as though it consisted of an unlimited number of individuals. This may be done, because the distribution of variates fully defines the variable and we may imagine any completely defined phenomenon to be repeated without end. The distribution of cases in the ideal series will then be such that the relative frequency of a measurement X, compared with the total number of measurements, may be expressed by the algebraical function $f(X)$. When, for instance, 20 observa-

Digitized by Google

tions have been made, and the measurement X_1 occurs 3 times, X_2 occurs 4 times, we have

$$f(X_1) = \tfrac{3}{20}, \quad f(X_2) = \tfrac{4}{20}.$$

Thus, from our actual observations we obtain a table of measurements from which may be derived an algebraical law representing this function with greater or less accuracy.

Two questions, therefore, arise at the beginning of our investigation — the one, how to determine this algebraical function; the other, how far the algebraical function thus obtained may be assumed to correspond to the function representing the unlimited series.

According to our definition all the measurements are to belong to the same class. It is obvious that the members of one and the same class can not vary more than a certain limited amount, which is determined by the definition of the class. If the variations contained in the class had an unlimited range, the class itself would be unlimited. In other words, in no case do variations occur of such size that the measurement could no longer belong to the class in question. Therefore variates must always remain within certain finite limits, or the frequency of variates beyond these limits is zero.

The character of the function expressing the distribution of variates may be determined by arranging the measurements in order, beginning with the lowest and proceeding to the highest, or *vice versa*. The measurements may be recorded with the greatest possible accuracy, which may be made so great that the observations may be given in form of a list of measurements of which each occurs only once, or at most a few times, and which will be crowded where the frequencies of the measured values are great, and far apart where the frequencies are low. A clearer impression of the character of the distribution may be obtained by counting the number of measurements that occur within convenient intervals. For instance, the statures of 905 nine-year-old boys in Toronto have been taken. By recording the number of cases that occur between groups of 10 mm., the following series results:

mm.	No. of Cases.	mm.	No. of Cases.	mm.	No. of Cases.
		1095–1105	1	1145–1155	2
1055–1065	1	1105–1115	2	1155–1165	9
1065–1075	—	1115–1125	3	1165–1175	17
1075–1085	2	1125–1135	3	1175–1185	27
1085–1095	—	1135–1145	3	1185–1195	21

Digitized by Google

mm.	No. of Cases.	mm.	No. of Cases.	mm.	No. of Cases.
1195–1205	31	1295–1305	65	1395–1405	2
1205–1215	52	1305–1315	41	1405–1415	2
1215–1225	43	1315–1325	48	1415–1425	1
1225–1235	50	1325–1335	35	1425–1435	1
1235–1245	54	1335–1345	27	1435–1445	—
1245–1255	67	1345–1355	19	1445–1455	—
1255–1265	63	1355–1365	16	1455–1465	1
1265–1275	73	1365–1375	5	1465–1475	—
1275–1285	62	1375–1385	9	1475–1485	1
1285–1295	43	1385–1395	2	1485–1495	—
				1495–1505	1

This distribution may be presented graphically by recording the intervals on a horizontal line and the frequencies as areas of rectangles erected over the intervals. This figure will be clearer, although not quite so accurate, if we simply mark by a point the middle of the upper limit of each rectangle, and connect these points by straight lines. If we consider the interval as a unit, the height of the rectangle will be equal to the recorded frequency.

The algebraical function to be found will then be that function which represents as nearly as possible all the numerical values of the observed relative frequencies. If we call these relative frequencies y and the measurements X, then $y = f(X)$. We may determine the total frequency of cases that occur up to various values of X; and we have

up to the lowest value X_1 the relative frequency	$y_1 = z_1$
" " next value X_2 " " "	$y_1 + y_2 = z_2$
" " third value X_3 " " "	$y_1 + y_2 + y_3 = z_3$
. . . . .	. . . . .
up to the nth value X_n the number	$y_1 + y_2 + \cdots + y_n = z_n$.

Then we must determine the number z for any value of X. This problem may be approached from two distinct points of view.

§ 3

Supposing that the values of z have been observed for a series of points, $X_1, X_2 \cdots, X_p$, we can represent the series of z between these points adequately, if we can find an algebraical function which, for the values $X_1 X_2 \cdots X_p$ assumes the values $z_1 z_2 \cdots z_p$. This condition may be fulfilled by many functions. A simple function that meets the requirements is

Digitized by Google

$$(1)\quad z = \frac{(X-X_2)(X-X_3)\cdots(X-X_p)}{(X_1-X_2)(X_1-X_3)\cdots(X_1-X_p)} z_1$$
$$+ \frac{(X-X_1)(X-X_3)\cdots(X-X_p)}{(X_2-X_1)(X_2-X_3)\cdots(X_2-X_p)} z_2 + \cdots$$
$$\cdots + \frac{(X-X_1)(X-X_2)\cdots(X-X_{r-1})(X-X_{r+1})\cdots(X-X_p)}{(X_r-X_1)(X_r-X_2)\cdots(X_r-X_{r-1})(X_r-X_{r+1})\cdots(X_r-X_p)} z_r + \cdots$$
$$\cdots + \frac{(X-X_1)(X-X_2)\cdots(X-X_{p-1})}{(X_p-X_1)(X_p-X_2)\cdots(X_p-X_{p-1})} z_p,$$

because in this equation, for $X = X_r$, all the coefficients, with the exception of the one not containing $(X - X_r)$, disappear, while the one that does not disappear will equal 1.

We assume in this formula, that in the region between the known points the function changes continuously and that no periodical or irregular features of the law of distribution occur in the intermediate regions.

The equation given before takes a simpler form if we compare any two succeeding interpolations for $2p$ and $2p + 1$ points. For a term with the factor z_r, we have, for the $2p$ points, from X_{-p+1} to X_p,

$$\frac{(X-X_{-p+1})(X-X_{-p+2})\cdots(X-X_{r-1})(X-X_{r+1})\cdots(X-X_p)}{(X_r-X_{-p+1})(X_r-X_{-p+2})\cdots(X_r-X_{r-1})(X_r-X_{r+1})\cdots(X_r-X_p)} z_r,$$

and for $2p + 1$ points from X_{-p} to X_p

$$\frac{(X-X_{-p})(X-X_{-p+1})\cdots(X-X_{r-1})(X-X_{r+1})\cdots(X-X_p)}{(X_r-X_{-p})(X_r-X_{-p+1})\cdots(X_r-X_{r-1})(X_r-X_{r+1})\cdots(X_r-X_p)} z_r,$$

for the $2p + 1$ points for X_{-p+1} to X_{p+1}

$$\frac{(X-X_{-p+1})(X-X_{-p+2})\cdots(X-X_{r-1})(X-X_{r+1})\cdots(X-X_{p+1})}{(X_r-X_{-p+1})(X_r-X_{-p+2})\cdots(X_r-X_{r-1})(X_r-X_{r+1})\cdots(X_r-X_{p+1})} z_r.$$

It follows that, for the $2p + 1$st interpolation, the following amounts must be added to the $2p$th interpolation; for points from X_{-p} to X_p

$$\frac{(X-X_{-p+1})(X-X_{-p+2})\cdots(X-X_p)}{(X_r-X_{-p})(X_r-X_{-p+1})\cdots(X_r-X_{r-1})(X_r-X_{r+1})\cdots(X_r-X_p)} z_r;$$

for points from X_{-p+1} to X_{p+1}

$$\frac{(X-X_{-p+1})(X-X_{-p+2})\cdots(X-X_p)}{(X_r-X_{-p+1})(X_r-X_{-p+2})\cdots(X_r-X_{r-1})(X_r-X_{r+1})\cdots(X_r-X_{p+1})} z_r.$$

Digitized by Google

If we take the values of X equidistant, their distances as units, if furthermore we assume the point X_0 as zero and call the varying values x, we find for the last two values

$$(2) \qquad \frac{(x+p-1)(x+p-2)\cdots(x-p)}{(r+p)(r+p-1)\cdots 1\cdot(-1)(-2)\cdots(r-p)}z_r$$

$$= (-1)^{r-p}\frac{2p(2p-1)\cdots(r+p+1)}{1\cdot 2\cdot 3\cdots(r-p)}z_r\frac{(x+p-1)(x+p-2)\cdots(x-p)}{1\cdot 2\cdot 3\cdots(2p)}$$

and

$$(2^*) \qquad \frac{(x+p-1)(x+p-2)\cdots(x-p)}{(r+p-1)(r+p-2)\cdots 2\cdot 1\cdot(-1)(-2)\cdots(r-p-1)}z_r$$

$$=(-1^{r-p-1})\frac{2p(2p-1)\cdots(r+p)}{1\cdot 2\cdot 3\cdots(r-p-1)}z_r\frac{(x+p-1)(x+p-2)\cdots(x-p)}{1\cdot 2\cdot 3\cdots 2p}.$$

It will be seen that the members for the points x_{-p} and x_{p+1} also agree with these forms. We have, therefore, if we call U_{2p} the value of the function for the interpolation of $2p$ points, from x_{-p+1} to x_p, U'_{2p+1} the interpolation for the $2p+1$ points from x_{-p} to x_p and U''_{2p+1} for the points x_{-p+1} to x_{p+1}, and if we designate by Σ the sum of the products (2) and (2*) for all values of r:

$$U'_{2p+1} = U_{2p}$$

$$+\frac{(x+p-1)(x+p-2)\cdots(x-p)}{1\cdot 2\cdot 3\cdots 2p}\Sigma(-1)^{r-p}\frac{2p(2p-1)\cdots(r+p+1)}{1\cdot 2\cdot 3\cdots(r-p)}z_r.$$

$$U''_{2p+1} = U_{2p}$$

$$+\frac{(x+p-1)(x+p-2)\cdots(x-p)}{1\cdot 2\cdot 3\cdots 2p}\Sigma(-1)^{r-p-1}\frac{2p(2p-1)\cdots(r+p)}{1\cdot 2\cdot 3\cdots(r-p-1)}z_r.$$

The sums in these two terms are the $2p$th differences between the values from z_{-p} to z_p, and from z_{-p+1} to z_{p+1}. This can be shown as follows: If the rth differences are expressed by the form

$$\Delta^r_{-p} = a_0 z_{-p} + a_1 z_{-p+1} + \cdots + a_{r+1} z_{-p+r+1},$$

then

$$\Delta^r_{-p+1} = a_0 z_{-p+1} + a_1 z_{-p+2} + \cdots + a_{r+1} z_{-p+r+2}.$$

Their difference

$$\Delta^{r+1}_{-p} = a_0 z_{-p} + (a_1 - a_0) z_{-p+1} + (a_2 - a_1) z_{-p+2} + \cdots$$
$$\cdots + (a_{r+1} - a_r) z_{-p+r+1} - a_{r+1} z_{-p+r+2}.$$

Digitized by Google

Since for Δ'_{-p} $a_0 = 1$, it follows at once that the coefficients a_1 decrease by units and are, therefore, for the rth difference $-r$. The coefficients a_2 increase by the amounts $a_1 = r$ and are, therefore, for the rth difference $r(r-1)/1 \cdot 2$. In short, the coefficients are the sums of arithmetical progressions of increasing order arising from whole numbers, but with alternating signs.

$$a_0 = 1,$$

$$a_1 = -r,$$

$$a_2 = +\frac{r(r-1)}{1 \cdot 2},$$

$$a_3 = -\frac{r(r-1)(r-2)}{1 \cdot 2 \cdot 3}, \text{ etc.}$$

If we write the differences between the values z in the following form:

$$\begin{array}{llll}
z_{-p} & & & \\
 & \Delta'_{-p} & & \\
z_{-p+1} & & \Delta''_{-p} & \\
 & \Delta'_{-p+1} & & \Delta'''_{-p} \\
z_{-2} & & \Delta''_{-p+1} & \\
 & \Delta'_{-2} & & \Delta'''_{-p+1} \\
z_{-1} & & \Delta''_{-2} & \\
 & \Delta'_{-1} & & \Delta'''_{-2} \\
z_0 & & \Delta''_{-1} & \\
 & \Delta'_0 & & \Delta'''_{-1} \\
z_1 & & \Delta''_0 & \\
 & \Delta'_1 & & \\
z_2 & & & \Delta'''_{p-4} \\
 & \Delta'_{p-2} & \Delta''_{p-3} & \\
z_{p-1} & & & \Delta'''_{p-3} \\
 & \Delta'_{p-1} & \Delta''_{p-2} & \\
z_p & & &
\end{array}$$

we find

$$U'_{2p+1} = U_{2p} + \frac{(x+p-1)(x+p-2)\cdots(x-p)}{1 \cdot 2 \cdot 3 \cdots 2p} \Delta^{2p}_{-p},$$

$$U''_{2p+1} = U_{2p} + \frac{(x+p-1)(x-p-2)\cdots(x-p)}{1 \cdot 2 \cdot 3 \cdots 2p} \Delta^{2p}_{-p+1}.$$

In the same way it can be shown that

$$U_{2p+2} = U'_{2p+1} + \frac{(x+p)(x+p-1)\cdots(x-p)}{1 \cdot 2 \cdot 3 \cdots (2p+1)} \Delta^{2p+1}_{-p};$$

and also

$$U_{2p+2} = U''_{2p+1} + \frac{(x+p-1)(x+p-2)\cdots(x-p-1)}{1 \cdot 2 \cdot 3 \cdots (2p+1)} \Delta^{2p+1}_{-p}.$$

Therefore

Digitized by Google

$$U_{2p+z} = \frac{U'_{2p+1} + U''_{2p+1}}{2} + \frac{(x+p-1)(x+p-2)\cdots(x-p)(x-\frac{1}{2})}{1\cdot 2\cdot 3\cdots(2p+1)}\Delta^{2p+1}_{-p}$$

$$= U_{2p} + \frac{(x+p-1)(x+p-2)\cdots(x-p)}{1\cdot 2\cdot 3\cdots 2p}\,\frac{\Delta^{2p}_{-p} + \Delta^{2p}_{-p+1}}{2}$$

$$+ \frac{(x+p-1)(x+p-2)\cdots(x-p)(x-\frac{1}{2})}{1\cdot 2\cdot 3\cdots(2p+1)}\Delta^{2p+1}_{-p}.$$

By introducing for p successively values from 0 upward, we find

$$(3)\; z = \frac{z_1+z_0}{2} + \frac{x-\frac{1}{2}}{1}\Delta'_0 + \frac{x(x-1)}{1\cdot 2}\,\frac{\Delta''_0+\Delta''_{-1}}{2} + \frac{x(x-1)(x-\frac{1}{2})}{1\cdot 2\cdot 3}\Delta'''_{-1}$$

$$+ \frac{(x+1)x(x-1)(x-2)}{1\cdot 2\cdot 3\cdot 4}\,\frac{\Delta^{IV}_{-1}+\Delta^{IV}_{-2}}{2}$$

$$+ \frac{(x+1)x(x-1)(x-2)(x-\frac{1}{2})}{1\cdot 2\cdot 3\cdot 4\cdot 5}\Delta^{V}_{-2} + \cdots.$$

Or

$$z = z_0 + x\Delta'_0 + \frac{x(x-1)}{4}(\Delta''_0+\Delta''_{-1}) + \frac{x(x-1)(x-\frac{1}{2})}{6}\Delta'''_{-1}$$

$$\frac{(x+1)x(x-1)(x-2)}{48}(\Delta^{IV}_{-1}+\Delta^{IV}_{-2}) + \cdots.$$

For various values of x the coefficients of interpolation are as follows:[1]

x	$\frac{x(x-1)}{4}$	$\frac{x(x-1)(x-\frac{1}{2})}{6}$	$\frac{(x+1)x(x-1)(x-2)}{48}$	x
0.00	−0.00000−	+0.0000−	+0.0000+	1.00
0.05	−0.01188−	+0.0036−	+0.0020+	0.95
0.10	−0.02250−	+0.0060−	+0.0039+	0.90
0.15	−0.03188−	+0.0074−	+0.0057+	0.85
0.20	−0.04000−	+0.0080−	+0.0072+	0.80
0.25	−0.04688−	+0.0078−	+0.0085+	0.75
0.30	−0.05250−	+0.0070−	+0.0097+	0.70
0.35	−0.05688−	+0.0057−	+0.0106+	0.65
0.40	−0.06000−	+0.0040−	+0.0112+	0.60
0.45	−0.06188−	+0.0021−	+0.0116+	0.55
0.50	−0.06250−	+0.0000−	+0.0117+	0.50

§ 4

In the method of interpolation described in § 3, we determine intermediate values of the function by assuming that a number of observed points between certain definite limits defines the function

[1] In this table the arguments from 0.0 to 0.5 are given on the left-hand side and are to be used with the sign on the left hand of the columns. Those from 0.5 to 1.0 are given on the right-hand side and are to be used with the corresponding sign.

Digitized by Google

adequately within these limits, and that the approach to the true distribution will be best in the middle region of the selected interval.

This method is open to the objection, that not all the observations are given equal weight, those near the middle region appearing with great weight, while those beyond the limits of the range of interpolation are entirely neglected.

It is possible to overcome this objection by demanding that each observation shall be given equal weight. This can be done by determining the averages of the powers of the variates. We may determine the average frequency of the variates, the average value of the variates, the averages of their squares, cubes, fourth powers, etc. If the function representing the unlimited series is determined, then the averages of these powers are also determined algebraically, and the averages of the observed limited series will approach the theoretical values more or less accurately. Therefore, if a function can be determined, whose average powers correspond to the calculated average powers of the observed series, it would seem that the problem can be solved in a more satisfactory manner than can be done by interpolation.

If we designate the measurements again by $X_1, X_2, \cdots, X_n$, we have to determine the averages

$$\frac{X_1^r + X_2^r + \cdots X_n^r}{n}.$$

Then we must find the function $f(X)$, for which the average of the infinitely large number of values $X^r f(X)$ corresponds to the observed averages. Or, if we indicate the process of averaging the infinitely long series by brackets,

$$\frac{X_1^r + X_2^r + \cdots + X_n^r}{n} = [X^r f(X)].$$

In applying this method we assume that any two functions which have the same average values of their powers are the same. This point requires, however, further investigation.

We will assume again that between certain limits, X_1 and X_2, the function can be represented by a series

$$f(X) = c_1 + c_2 X + c_3 X^2 + c_4 X^3 + \cdots c_{n+1} X^n.$$

Then the total number of cases between the limits X_1 and X_2 will be[1]

[1] For readers not familiar with the elements of calculus, the following explanation will perhaps be sufficient. If a certain function $f(x)$ is given, its total values

Digitized by Google

$$\int_{X_1}^{X_2} f(X)dx = c_1(X_2 - X_1) + c_2\frac{(X_2^2 - X_1^2)}{2} + c_3\frac{X_2^3 - X_1^3}{3} + \cdots$$

$$\cdots + c_{n+1}\frac{X_2^{n+1} - X_1^{n+1}}{n+1},$$

between two points x_2 and $x_2 + dx$ may be measured by the area bounded by the ordinates $f(x_2)$, $f(x_2 + dx)$, the line x_2, $x_2 + dx$ and the corresponding segment of the curve. When dx is taken very small, this area will be very nearly $f(x_2)dx$. The area bounded by the curve, the terminal ordinates $f(x_1)$, $f(x_2)$, and the line x_1x_2, will be equal to the sum of all the values $f(x)dx$ between the limits x_1 and x_2, which may be written

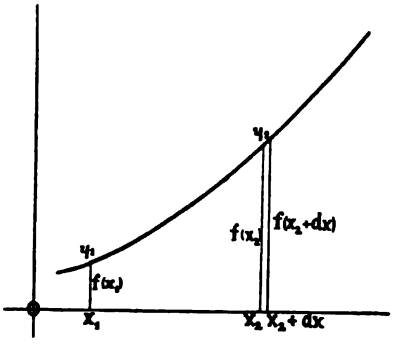


FIG. 1.

$$\int_{x1}^{x_2} f(x)dx.$$

We may also consider the area thus bounded between the points x_1 and a variable point x as a function of x, $\phi(x)$. Then we can write

$$\phi(x + dx) - \phi(x) = f(x)dx$$

or

$$\frac{\phi(x + dx) - \phi(x)}{dx} = f(x).$$

If we assume, for instance, $\phi(x) = x^n$, when n is an integer we find

$$\frac{(x + dx)^n - x^n}{dx} = f(x)$$

and expanded

$$n \cdot x^{n-1} + \frac{n(n-1)}{1 \cdot 2}x^{n-2}dx + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}x^{n-3}dx^2 + \cdots = f(x).$$

Since we can make dx as small as we please,

$$nx^{n-1} = f(x).$$

Conversely we may conclude, that when the terminal coordinate is x^n, the area from a certain initial point must be

$$\int_o^x x^n dx = \frac{x^{n+1}}{n+1},$$

as may be shown by substitution in the preceding formula. If we calculate this series first up to the point x_2, then to x_1, we obtain the area bounded by $x_1x_2y_2y_1$ and by the curve,

$$\int_{x_1}^{x_2} x^n dx = \frac{x_2^{n+1} - x_1^{n+1}}{n+1}.$$

This is the formula applied in the text.

Digitized by Google

and for the pth power

$$\int_{X_1}^{X_2} x^p f(X)dx = c_1 \frac{X_2^{p+1} - X_1^{p+1}}{p+1} + c_2 \frac{X_2^{p+2} - X_1^{p+2}}{p+2} + \cdots$$

$$\cdots + c_{n+1} \frac{X_2^{n+p+1} - X_1^{n+p+1}}{n+p+1}.$$

If we determine the powers with X_1 as initial point,[1] and designate these powers by $a_0^0, a_1^1, a_2^2, \cdots$, we find

$$a_p^p = c_1 \frac{X_2^{p+1}}{p+1} + c_2 \frac{X_2^{p+2}}{p+2} + \cdots c_{n+1} \frac{X_2^{n+p+1}}{n+p+1}. \tag{4}$$

It follows at once that when in two functions the values of a are the same between the limits X_1 and X_2, and the functions between these limits can be expressed by a limited series containing only powers of x, the functions between these limits are also the same, because the constants $c_1, c_2 \cdots c_n$ will be the same. In that case we have the two equations

$$a_p^p = c_1 \frac{X_2^{p+1}}{p+1} + c_2 \frac{X_2^{p+2}}{p+2} + \cdots c_n \frac{X_2^{n+p+1}}{n+p+1},$$

$$a_p^p = c_1' \frac{X_2^{p+1}}{p+1} + c_2' \frac{X_2^{p+2}}{p+2} + \cdots c_n' \frac{X_2^{n+p+1}}{n+p+1},$$

from which follows

$$0 = (c_1 - c_1') \frac{X_2^{p+1}}{p+1} + (c_2 - c_2') \frac{X_2^{p+2}}{p+2} + \cdots (c_n - c_n') \frac{X_2^{n+p+1}}{n+p+1}.$$

This can be true only when the series of

$$c = c'.$$

It appears also that the constants of the function may be determined from (4) by successive elimination.

[1] See pp. 26, *et seq.*

Digitized by Google

II. COMPARISON BETWEEN LIMITED SERIES OF OBSERVATIONS AND THE UNLIMITED SERIES OF VARIABLES

A. Properties of Averages

§ 5

In the method outlined in § 4 it is presupposed that we know the exact values of the averages of the powers of X. However, in a number of limited series we do not expect to obtain uniformly the same relative frequency for each variate, because there will be accidental differences due to the limited number of cases, and a strict correspondence between the distribution of the limited series and of the unlimited series does not exist. For this reason we can not find the exact values of the averages of the powers of X, but only approximations which will vary according to the accidental peculiarities of each series.

The question arises, therefore, whether we can determine the characteristics of the distribution of the averages of powers of X. In investigating this problem we shall use the method suggested in § 4 and try to determine the averages of the powers of those values which express the distribution of the limited averages around the general average. We have called the general average of the pth powers of our unlimited series a_p^p. The corresponding special averages of limited series each containing n observations may be called $a_p'^p$. Then we shall determine the values of

$$[(a_p'^p - a_p^p)^r].$$

We shall direct particular attention to the values of

$$[(a' - a)^r],$$

the powers of the differences between the special averages and the general average of the function.

Before we take up this subject, it will be well to discuss a few general properties of averages.

The average of the sum of two variables is equal to the sum of their averages.

Digitized by Google

$$[y+z] = \frac{(y_1+z_1)+(y_2+z_2)+\cdots+(y_N+z_N)}{N}$$

$$= \frac{y_1+y_2+\cdots+y_N}{N} + \frac{z_1+z_2+\cdots+z_N}{N}.$$

(5) $[y+z] = [y] + [z]$.

The averages of the product of two independent variables is equal to the product of their averages.

$$[yz] = \frac{y_1z_1 + y_2z_2 + \cdots + y_Nz_N}{N}.$$

The members of this sum may be so grouped that all the values y_r that have the same numerical value are grouped together. Then each value y_r will appear as the factor of a sum,

$$z_r' + z_r'' + \cdots z_r^M,$$

which will have the average value of $m[z]$. Thus every value of y_r appears as the factor of a certain value $m[z]$ where m equals the number of occurrences of the respective value y_r. We find, therefore,

$$[yz] = [z]\,\frac{m_1y_1 + m_2y_2 + \cdots + m_Ny_N}{N},$$

(6) $$[yz] = [y]\,[z].$$

We may now take up the discussion of the values $[(a'-a)^r]$; and we will begin with the consideration of $[(a'-a)]$.

For a special series of n observations, the special average

$$a' = \frac{X_1' + X_2' + \cdots + X_n'}{n},$$

therefore

$$[(a'-a)] = \left[\frac{X_1'-a}{n} + \frac{X_2'-a}{n} + \cdots + \frac{X_n'-a}{n}\right].$$

We will write

$$X - a = x.$$

$$[(a'-a)] = \left[\frac{x_1'}{n} + \frac{x_2'}{n} + \cdots + \frac{x_n'}{n}\right].$$

Since according to definition $a = [X]$,

$$[x] = [X-a] = 0,$$

therefore

$$[(a'-a)] = 0.$$

Digitized by Google

We will proceed in the same manner to evaluate $[(a' - a)^2]$. As before, we can write

$$[(a' - a)^2] = \left[\left(\frac{x_1'}{n} + \frac{x_2'}{n} + \cdots + \frac{x_n'}{n}\right)^2\right]$$

$$= \frac{[x_1'^2]}{n^2} + \frac{[x_2'^2]}{n^2} + \cdots + \frac{[x_n'^2]}{n^2}$$

$$+ \frac{[x_1'x_2']}{n^2} + \frac{[x_1'x_3']}{n^2} + \cdots$$

$$+ \frac{[x_2'x_1']}{n^2} + \frac{[x_2'x_3']}{n^2} + \cdots.$$

The general average of the expression x^2 depends solely on the character of the function which determines the distribution of X, viz., x, and we may write

$$[x^2] = \sigma^2.$$

Then each of the n expressions of the form

$$\frac{[x^2]}{n^2} = \frac{\sigma^2}{n^2},$$

hence, their sum total

$$\frac{[x_1'^2]}{n^2} + \frac{[x_2'^2]}{n^2} + \cdots \frac{[x_n'^2]}{n^2} = \frac{\sigma^2}{n}.$$

The second group of terms of the type

$$\frac{[x_r x_q]}{n^2} = \frac{[x_r][x_q]}{n^2}.$$

Since each of these factors averages zero, all their products will be zero. Thus we find

$$[(a' - a)^2] = \frac{\sigma^2}{n},$$

or

$$\sqrt{[(a' - a)^2]} = \pm \frac{\sigma}{\sqrt{n}}.$$

We will also estimate the mean of the cubes and fourth powers of the differences between the special averages and the general average.

$$(a' - a)^3 = \left(\frac{x_1'}{n} + \frac{x_2'}{n} + \cdots + \frac{x_n'}{n}\right)^3.$$

Digitized by Google

In expanding this expression we obtain n terms of the form x'^3/n^3. All the other terms will contain squares of x multiplied by first powers of x

$$[x_r^2 x_q] = [x_r^2][x_q] = 0.$$

Therefore,

$$[(a' - a)^3] = \frac{[x^3]}{n^2}.$$

In the same way

$$(a' - a)^4 = \left(\frac{x_1'}{n} + \frac{x_2'}{n} + \frac{x_3'}{n} + \cdots + \frac{x_n'}{n}\right)^4.$$

Here again all those products in the expansion of the polynomial which contain a first power of x will be zero, hence

$$[(a' - a)^4] = \frac{[x^4]}{n^3} + 6\frac{n(n-1)[x^2]^2}{n^4 \cdot 1 \cdot 2}.$$

If we call

$$[x^3] = \sigma_3^3,$$

and

$$[x^4] = \sigma_4^4,$$

we have

$$(7) \quad \begin{cases} [(a' - a)] = 0, \\ [(a' - a)^2] = \dfrac{\sigma^2}{n}, \\ [(a' - a)^3] = \dfrac{\sigma_3^3}{n^2}, \\ [(a' - a)^4] = \dfrac{\sigma_4^4}{n^3} + 3\dfrac{(n-1)\sigma^4}{n^3}. \end{cases}$$

We will designate the values of

$$\frac{\sigma}{\sqrt{n}} = \epsilon, \qquad \frac{\sigma_3}{\sqrt{n}} = \epsilon_3, \qquad \frac{\sigma_4}{\sqrt{n}} = \epsilon_4.$$

Then we may write

$$(7^*) \quad \begin{cases} [(a' - a)] = 0, \\ [(a' - a)^2] = \epsilon^2, \\ [(a' - a)^3] = \dfrac{\epsilon_3^3}{\sqrt{n}}, \\ [(a' - a)^4] = 3\epsilon^4 + \dfrac{\epsilon_4^4 - 3\epsilon^4}{n}. \end{cases}$$

Digitized by Google

These averages, which may be used to determine the characteristics of the distribution of a', express, at the same time, the mean values of the powers of the differences $a' - a$ that may be expected in a series of n observations and may, therefore, be called expected errors.

We will now proceed to evaluate the higher powers of $(a' - a)$. It is convenient to treat the even powers and the odd powers separately. We will first determine

$$[(a' - a)^{2p}] = \left[\left(\frac{x_1'}{n} + \frac{x_2'}{n} + \cdots + \frac{x_n'}{n}\right)^{2p}\right].$$

If we expand this polynomial term we can write

$$[(a' - a)^{2p}] = \frac{1}{n^{2p}}\left[\Sigma \frac{2p(2p-1)\cdots 2\cdot 1}{(1\cdot 2\cdots r_1)(1\cdot 2\cdots r_2)\cdots(1\cdot 2\cdots r_u)} \times (x_{i_1}^{r_1} x_{i_2}^{r_2} \cdots x_{i_{u-1}}^{r_{u-1}} x_{i_u}^{2p-r_1-r_2-\cdots-r_{u-1}})\right].$$

Since all the values of x represent the same function, every

$$[x^r] = \sigma_r^r.$$

For any series containing only different values of

$$r_1, r_2 \cdots r_{u-1}, 2p - r_1 - r_2 - \cdots - r_{u-1},$$

among a total of n members of the polynomial expression, we find, therefore, $n(n-1)\cdots(n-u+1)$ different combinations which contain the same product

$$\sigma_{r_1}^{r_1}\sigma_{r_2}^{r_2}\sigma_{r_{u-1}}^{r_{u-1}} \cdots \sigma_{r_u}^{2p-r_1-r_2-\cdots-r_{u-1}}.$$

If in this series there are $l_1, l_2 \cdots l_v$ members for which the exponents are the same, their number will be

$$\frac{n(n-1)\cdots(n-u+1)}{(1\cdot 2\cdots l_1)(1\cdot 2\cdots l_2)\cdots(1\cdot 2\cdots l_v)}.$$

In our particular case, all those products which contain the first power of an x will be zero, because the average value of every x is zero. Therefore only those values need be considered for which $r > 1$. As long as $n - u + 1$ is greater than any value l_r, these products will be the greater, the greater u is. Since, however, x must not appear with an exponent less than 2, the maximum value for u will be p. In this case all the p exponents are the same, and we find, therefore, for the number of equal terms

$$\frac{n(n-1)\cdots(n-p+1)}{1\cdot 2\cdots p}.$$

Digitized by Google

The average value of each x^2 is σ^2; therefore, since all the values $r = 2$

$$\frac{1}{n^{2p}}\left[\Sigma \frac{2p(2p-1)\cdots(2p-r_1-r_2-\cdots-r_{u-1}+1)}{(1\cdot 2\cdots r_1)(1\cdot 2\cdots r_2)\cdots(1\cdot 2\cdots r_{u-1})}\right.$$

$$\left.\Sigma(x_{i_1}^{r_1}x_{i_2}^{r_2}\cdots x_{i_{u-1}}^{r_{u-1}}x_{i_u}^{2p-r_1-r_2-\cdots-r_{u-1}})\right]$$

$$= \frac{1}{n^{2p}}\left\{\frac{2p(2p-1)}{1\cdot 2}\frac{(2p-2)(2p-3)}{1\cdot 2}\cdots\frac{2\cdot 1}{1\cdot 2}\right\}\frac{n(n-1)\cdots(n-p+1)}{1\cdot 2\cdots p}\sigma^{2p}$$

$$= \frac{1}{n^{2p}}\{(2p-1)(2p-3)\cdots 3\cdot 1\}\{n(n-1)\cdots(n-p+1)\}\sigma^{2p},$$

and, if we introduce again

$$\frac{\sigma}{\sqrt{n}} = \epsilon$$

$$(8) \quad = \{(2p-1)(2p-3)\cdots 3\cdot 1\}\left\{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{p-1}{n}\right)\right\}\epsilon^{2p}$$

For any value $u < p$ the product of $n(n-1)\cdots(n-u+1)$ will contain a less number of members than occur when every member has the exponent 2; and the factors of our product containing the powers of ϵ, ϵ_3, ϵ_4, $\cdots$ will be divided by $\sqrt{n^{p-u}}$.

Now it can be shown that the values of ϵ, ϵ_3, ϵ_4, $\cdots$ must be nearly of the same order. According to the remark made on p. 5 the deviations in any class must be limited. If we designate their limit by L,

$$[x^p] < L^p,$$

$$\sqrt[p]{[x^p]} < L.$$

Therefore the smaller L, the more nearly will all the values σ be small as compared with n. If n is assumed sufficiently large, all the terms containing n in the denominator may be neglected, and thus we find

$$(9) \qquad [(a'-a)^{2p}] = \{(2p-1)(2p-3)\cdots 3\cdot 1\}\epsilon^{2p}.$$

The same consideration shows that for large values of n

$$(9^*) \qquad [(a'-a)^{2p+1}] = 0.$$

These approximations are sufficient as long as p is small when compared with n. For large values of p the approximation is not correct.

Digitized by Google

This series of averages of powers corresponds to the average powers of the exponential formula

$$y = \frac{1}{\epsilon\sqrt{2\pi}} e^{-\frac{(a'-a)^2}{2\epsilon^2}}. \tag{10}$$

We shall next determine the function which arises when we consider terms of the order $1/\sqrt{n}$.

For even powers $[(a'-a)^{2p}]$, values of this order originate through the combination of $(p-2)$ elements of the order x^2, one element of the order x^3 and one element of the order x. Since the average of the last of these averages equals zero, the values of $[x^{2p}]$ are not changed by introducing elements of the order $1/\sqrt{n}$.

For odd powers $[(a'-a)^{2p+1}]$, the values of the order $1/\sqrt{n}$ originate by the combination of $(p-1)$ elements of the order x^2, and one element of the order x^3.

$$[(a'-a)^{2p+1}]$$

$$= \frac{1}{n^{2p}} \frac{(2p+1)2p(2p-1)\cdots 2\cdot 1}{(1\cdot 2)(1\cdot 2)\cdots(1\cdot 2\cdot 3)} \cdot \frac{n(n-1)\cdots(n-p+1)}{1\cdot 2\cdots(p-1)} \sigma^{2p-2}\sigma_3^3,$$

$$= \frac{(2p+1)2p(2p-1)}{1\cdot 2\cdot 3\sqrt{n}} \{(2p-3)(2p-5)\cdots 3\cdot 1\}$$

$$\times \left\{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{p-1}{n}\right)\right\} \epsilon^{2p-2}\epsilon_3^3,$$

or by neglecting terms of orders higher than $1/\sqrt{n}$,

$$= p\frac{(2p+1)(2p-1)\cdots 7\cdot 5}{\sqrt{n}} \epsilon^{2p-2}\epsilon_3^3. \tag{11}$$

In order to determine the function corresponding to these average powers, we may write

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}(1 + bx + cx^3),$$

which must be its type, because the even moments give values corresponding to the exponential function. Then we have

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}(x + bx^2 + cx^4)dx = b\epsilon^2 + 3c\epsilon^4 = 0,$$

$$\int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}(x^3 + bx^4 + cx^6)dx = 3b\epsilon^4 + 15c\epsilon^6 = \frac{\epsilon_3^3}{\sqrt{n}},$$

Digitized by Google

$$c = \frac{\epsilon_3^3}{6\epsilon^6\sqrt{n}},$$

$$b = -\frac{\epsilon_3^3}{2\epsilon^4\sqrt{n}},$$

and

$$y = \frac{1}{\epsilon\sqrt{2\pi}} e^{-\frac{x^2}{2\epsilon^2}} \left\{ 1 - \frac{\epsilon_3^3}{2\epsilon^4\sqrt{n}} \left(x - \frac{x^3}{3\epsilon^2} \right) \right\}.$$

The odd moments of this function have the values

$$[x^{2p+1}] = -[\{(2p+1)(2p-1)\cdots 5\cdot 3\} - \{(2p+3)(2p+1)\cdots 7\cdot 5\}] \times \frac{\epsilon^{2p-2}\epsilon_3^3}{2\sqrt{n}} = p\frac{(2p+1)(2p-1)\cdots 7\cdot 5}{\sqrt{n}} \epsilon^{2p-2}\cdot\epsilon_3^3,$$

and therefore agree with the powers demanded in (11).

We will also consider the form that our function takes, provided the members of the order $1/n$ are considered. These leave the odd powers as found before. The first two terms of even powers $[(a'-a)^{2p}]$ are found by including in the expression (8) terms with the denominator n. Besides these, elements of the order $1/n$ originate through the combination of $(p-2)$ elements x^2, and one element x^4; and also through the combination of $(p-3)$ elements x^2, and two elements x^3. Thus we obtain

$$\begin{aligned}[(a'-a)^{2p}] &= \{(2p-1)(2p-3)\cdots 3\cdot 1\}\epsilon^{2p} \\ &\quad - \frac{p(p-1)}{1\cdot 2n}\{(2p-1)(2p-3)\cdots 3\cdot 1\}\epsilon^{2p} \\ &\quad + \frac{p(p-1)(p-2)}{9n}\{(2p-1)(2p-3)\cdots 3\cdot 1\}\epsilon^{2p-6}\epsilon_3^6 \\ &\quad + \frac{2p(2p-1)(2p-2)(2p-3)}{1\cdot 2\cdot 3\cdot 4\cdot n} \\ &\qquad \times \{(2p-5)(2p-7)\cdots 3\cdot 1\}\epsilon^{2p-4}\epsilon_4^4 \\ &= \{(2p-1)(2p-3)\cdots 3\cdot 1\}\epsilon^{2p} \\ &\qquad \left\{ 1 + \frac{p(p-1)}{2n}\left(\frac{2(p-2)\epsilon_3^6}{9\epsilon^6} + \frac{\epsilon_4^4}{3\epsilon^4} - 1 \right) \right\}.\end{aligned}$$

The function giving these average odd and even powers has the form

$$y = \frac{1}{\epsilon\sqrt{2\pi}} e^{-\frac{x^2}{2\epsilon^2}} \left\{ 1 - \frac{\epsilon_3^3}{2\epsilon^4\sqrt{n}}\left(x - \frac{x^3}{3\epsilon^2} \right) + \frac{\epsilon_4^4 - 3\epsilon^4}{8n\epsilon^4}\left(1 - \frac{2x^2}{\epsilon^2} + \frac{x^4}{3\cdot 1\epsilon^4} \right) + \frac{5}{24}\frac{\epsilon_3^6}{n\epsilon^6}\left(1 - \frac{3x^2}{\epsilon^2} + \frac{3x^4}{3\cdot 1\epsilon^4} - \frac{x^6}{5\cdot 3\cdot 1\epsilon^6} \right) \right\}. \tag{12}$$

Digitized by Google

As soon as p is large when compared with n, the approximations are no longer satisfactory. It can, however, be shown that differences between averages of higher powers do not materially alter the function expressed by these average powers.

Supposing we have two functions of the forms

$$\frac{1}{\epsilon\sqrt{2\pi}} e^{-\frac{x^2}{2\epsilon^2}}(a_0 + a_1 x + a_2 x^2 + \cdots + a_{2m} x^{2m})$$

and

$$\frac{1}{\epsilon\sqrt{2\pi}} e^{-\frac{x^2}{2\epsilon^2}}(a_0' + a_1' x + a_2' x^2 + \cdots + a_{2m}' x^{2m}).$$

We will assume that their $2m-1$ first moments are the same, while the higher moments differ. If we call

$$a - a' = d$$

and the difference between the $2m$th moments s

$$d_0 + 1 \cdot d_2\epsilon^2 + 3 \cdot 1 d_4\epsilon^4 + 5 \cdot 3 \cdot 1 \cdot d_6\epsilon^6 + \cdots$$
$$\cdots + \{(2m-1)(2m-3)\cdots 3 \cdot 1\} d_{2m}\epsilon^{2m} = 0,$$

$$d_0 + 3d_2\epsilon^2 + 5 \cdot 3 d_4\epsilon^4 + 7 \cdot 5 \cdot 3 d_6\epsilon^6$$
$$+ \cdots + \{(2m+1)(2m-1)\cdots 3\} d_{2m}\epsilon^{2m} = 0,$$

. . . . . .

$$d_0 + (2m-1)d_2\epsilon^2 + (2m+1)(2m-1)d_4\epsilon^4 + (2m+3)(2m+1)(2m-1)d_6\epsilon^6$$
$$+ \cdots + \{(4m-3)(4m-5)\cdots(2m-1)\} d_{2m}\epsilon^{2m} = 0,$$

$$d_0 + (2m+1)d_2\epsilon^2 + (2m+3)(2m+1)d_4\epsilon^4 + (2m+5)(2m+3)(2m+1)d_6\epsilon^6$$
$$+ \cdots + \{(4m-1)(4m-3)\cdots(2m+1)\} d_{2m}\epsilon^{2m}$$
$$= \frac{s}{\epsilon^{2m}\{(2m-1)(2m-3)\cdots 3 \cdot 1\}}.$$

By consecutive elimination of the $r-1$ first elements, we find for the tth element of the first remaining equation the coefficient

$$\{2t(2t-2)(2t-4)\cdots(2t-2r-1)\}\{(2t-1)(2t-3)\cdots(2r-1)\},$$

and by subsequent elimination of the r' last elements we find for the tth element of the first remaining equation the coefficient

$$\{2t(2t-2)(2t-4)\cdots(2t-2r-1)\}$$
$$\times \{(2m-2t)(2m-2t-2)\cdots(2m-2r'-2t+2)\}$$
$$\times \{(2t-1)(2t-3)\cdots(2r-1)\},$$

and by taking

$$t = r \quad \text{and} \quad r' = 2m - r,$$

Digitized by Google

we have for the coefficient of $d_{2r}\epsilon^{2r}$

$$\{2r(2r-2)\cdots 4\cdot 2\}\{(2m-2r)(2m-2r-2)\cdots 4\cdot 2\}.$$

Since in these consecutive operations the values of the right-hand side are zero, except the last one, we need consider only the last one, which is multiplied in the process of elimination by $\{(2m-1)(2m-3)\cdots(2r-1)\}$. Thus we find

$$d_{2r} = \frac{s(2r-1)}{\epsilon^{2m+2r}\{2r(2r-1)\cdots 2.1\}\{(2m-2r)(2m-2r-2)\cdots 4\cdot 2\}}$$

Since m is assumed to be a large value, the values d—which are the changes in a due to the difference in the higher moment—are very small, unless s is a large and ϵ a very small value. If we take into consideration the differences between still higher moments, it can be shown in the same way that their influences are small.

§ 6

Since the particular average a' is an imperfect representation of the average value of the function, the differences $a'-a$, according to the terminology of § 1, may be called errors of observation. According to (7*) we have

$$[(a'-a)^2] = \epsilon^2 = \frac{\sigma^2}{n},$$

where

$$\sigma^2 = [(X-a)^2].$$

In this equation σ^2 is determined by the whole infinitely long series of observations. In any actual series of n observations σ^2 is not known, but only the approximate value σ'^2 derived by averaging the n values $(X'-a')^2$. It is, therefore, necessary to establish the relation between σ and σ'.

$$\sigma'^2 = \frac{(X_1'-a')^2+(X_2'-a')^2+\cdots+(X_n'-a')^2}{n},$$

$$= \frac{\{(X_1'-a)-(a'-a)\}^2+\{(X_2'-a)-(a'-a)\}^2+\cdots+\{(X_n'-a)-(a'-a)\}^2}{n},$$

$$= \frac{(X_1'-a)^2+(X_2'-a)^2+\cdots(X_n'-a)^2-2(a'-a)(X_1'-a+X_2'-a+\cdots+X_n'-a)+n(a'-a)^2}{n}.$$

Digitized by Google

Since

$$X_1' + X_2' + \cdots X_n' = na',$$

$$\sigma'^2 = \frac{(X_1' - a)^2 + (X_2' - a)^2 + \cdots + (X_n' - a)^2 - n(a' - a)^2}{n},$$

$$[(X - a)^2] = \sigma^2,$$

and

$$[(a' - a)^2] = \frac{\sigma^2}{n}, \tag{7}$$

therefore

$$\sigma'^2 = \sigma^2 \frac{n-1}{n},$$

and by substitution in (7*)

$$\epsilon = \sqrt{[(a - a')^2]} = \frac{\sigma'}{\sqrt{n-1}}. \tag{13}$$

This equation may be expressed in words as follows: The expected mean square error of the average is proportional to the mean square deviation and inversely proportional to the square root of the number of cases less one.

As an example the measurements of the proportions of length and breadth of head of twenty-five Indians of the interior of British Columbia may be given. The average obtained from the series is 83.4. The first column in the following table contains the measurements X; the second, their frequencies $F(X)$; the third, the values of $(X - a')$; the fourth, $(X - a')^2$; the fifth, these squares multiplied by their frequencies, and at the foot of the columns, the computation of the mean square deviation of the series.

$n = 25$; Average $= 83.4$.

X	$F(X)$	$X - a'$	$(X - a')^2$	$F(X)(X - a')^2$
77	1	−6.4	40.96	40.96
78	—	−5.4	29.16	
79	1	−4.4	19.36	19.36
80	2	−3.4	11.56	23.12
81	2	−2.4	5.76	11.52
82	3	−1.4	1.96	5.88
83	5	−0.4	0.16	0.80
84	3	+0.6	0.36	1.08
85	3	+1.6	2.56	7.68
86	1	+2.6	6.76	6.76
87	1	+3.6	12.96	12.96
88	1	+4.6	21.16	21.16
89	2	+5.6	31.36	62.72
				Total, 214.00

$$\sigma'^2 = \frac{\Sigma(X - a')^2}{n} = 8.56,$$

Digitized by Google

$$\sigma^2 = \frac{\Sigma(X - a')^2}{n-1} = 8.91,$$

$$\sigma = \sqrt{\frac{\Sigma(X' - a)^2}{n-1}} = \pm 2.98.$$

Mean square error $= \pm 2.98/4.9 = \pm 0.608$, that is to say, we may expect that, on the average, the square root of the mean square difference between the average of the typical series and our special series will have the value 0.608, either positive or negative.

In this manner the constant in (10) may be determined, and by means of a table, giving the values of the function

$$y = \frac{1}{\epsilon\sqrt{2\pi}} e^{\frac{-(a'-a)^2}{2\epsilon^2}},$$

the frequency of any particular value of a' may be determined.

If we do not make the approximation from which the exponential formula was derived, but go back to the formula (7) (p. 17), it appears that the values $\sigma_3, \sigma_4 \cdots$ must be determined. These values are not known, and their relation to $\sigma_3', \sigma_4' \cdots$ must be investigated. The method may be illustrated by the discussion of the value of σ_3'.

We proceed as before (p. 23), and write for every term

$$X' - a' = (X' - a) - (a' - a).$$

The cubic expansion of the average gives

$$\frac{1}{n}\{(X_1' - a')^3 + (X_2' - a')^3 + \cdots + (X_n' - a')^3\}$$

$$= \frac{1}{n}\{(X_1' - a)^3 + (X_2' - a)^3 + \cdots + (X_n' - a)^3\}$$

$$- \frac{3(a' - a)}{n}\{(X_1' - a)^2 + (X_2' - a)^2 + \cdots + (X_n' - a)^2\}$$

$$+ \frac{3(a' - a)^2}{n}\{(X_1' - a) + (X_2' - a) + \cdots + (X_n' - a\} - \frac{n(a' - a)^3}{n}.$$

The four terms on the right-hand side of this equation may be averaged singly

$$\left[\frac{1}{n}\{(X_1' - a)^3 + (X_2' - a)^3 + \cdots + (X_n' - a)^3\}\right] = \sigma_3^3.$$

For the second term we must determine the average of the expression of the form $(X_1' - a)^2(a' - a)$, whose factors are not entirely

Digitized by Google

independent of one another. If we arrange these so that all X_1 that have the same value are grouped together, we can write

$$a' = \frac{1}{n}(X_1 + X_2' + X_3' + \cdots + X_n')$$

$$a'' = \frac{1}{n}(X_1 + X_2'' + X_3'' + \cdots + X_n'')$$

. . . . . .

Since we have not made any conditions for $X_2, X_3, \cdots X_n$, the average of each of them will be a, and we have

$$[a'] = \frac{1}{n}[X_1 + (n-1)a] = a + \frac{(X_1 - a)}{n}.$$

Therefore for a certain X_1

$$[(X_1 - a)^2(a' - a)] = \frac{(X_1 - a)^3}{n},$$

and for all possible X

$$[(X_1' - a)^2(a' - a)] = \frac{\sigma_3^3}{n}.$$

In the third term the sum $[\{(X_1'-a)+(X_2'-a)+\cdots+X_n'-a)\}]$ for any particular a' equals $n(a' - a)$. Therefore the whole term will equal $3(a' - a)^3$. According to (7)

$$3[(a' - a)^3] = \frac{3\sigma_3^3}{n^2}.$$

Therefore

$$[(X' - a')^3] = \sigma_3^3 - \frac{3\sigma_3^3}{n} + \frac{3\sigma_3^3}{n^2} - \frac{\sigma_3^3}{n^2}$$

$$= \sigma_3^3\left(1 - \frac{3}{n} + \frac{2}{n^2}\right) = \frac{\sigma_3^3(n-1)(n-2)}{n^2}$$

$$\sigma_3'^3 = \sigma_3^3\frac{(n-1)(n-2)}{n^2}. \tag{14}$$

§ 7

It seems desirable to discuss at this point a few properties of the powers of a function. We will call, as heretofore,

$$a = [X],$$

and the deviations from the average

$$x = X - a.$$

Digitized by Google

It follows that

$$[x] = 0$$

and

$$[X^r] = [(a+x)^r].$$

$$(15)\quad \begin{cases} [X] = a, \\ [X^2] = a^2 + [x^2], \\ [X^3] = a^3 + 3a[x^2] + [x^3], \\ [X^4] = a^4 + 6a^2[x^2] + 4a[x^3] + [x^4], \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ [X^r] = a^r + \dfrac{r(r-1)}{1\cdot 2}a^{r-2}[x^2] \\ \qquad\qquad + \dfrac{r(r-1)(r-2)}{1\cdot 2\cdot 3}a^{r-3}[x^3] + \cdots + [x^r]. \end{cases}$$

In calculating the values of $[x^r]$ it is convenient to measure X from an arbitrary unit situated near a, which we will call c. We will also call

$$c - a = d$$

and

$$X = c + z;$$

then

$$c + z = a + x,$$

$$z = x - d,$$

$$[z^r] = [(x-d)^r],$$

and, as above,

$$(16)\quad \begin{cases} [z] = -d, \\ [z^2] = [x^2] + d^2, \\ [z^3] = [x^3] - 3d[x^2] - d^3, \\ [z^4] = [x^4] - 4d[x^3] + 6d^2[x^2] + d^4, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ [z^r] = [x^r] - rd[x^{r-1}] + \dfrac{r(r-1)}{1\cdot 2}d^2[x^{r-2}] + \cdots \\ \qquad\qquad \cdots + \dfrac{r(r-1)}{1\cdot 2}(-d)^{r-2}[x^2] + (-d)^r. \end{cases}$$

By applying formula (16) the calculation of $[x^r]$ may be much simplified, since we have to take the averages only of the powers of the small integers z, instead of the fraction x. For the table given in § 6 we may assume, for instance,

$$c = 83.$$

Digitized by Google

Then the original table may be arranged as follows:

X	$z = X - c$	$F(z)$	$zF(z)$	z^2	$z^2F(z)$
77	−6	1	−6	36	36
78	−5			25	
79	−4	1	−4	16	16
80	−3	2	−6	9	18
81	−2	2	−4	4	8
82	−1	3	−3	1	3
83	0	5		0	
84	+1	3	+3	1	3
85	+2	3	+6	4	12
86	+3	1	+3	9	9
87	+4	1	+4	16	16
88	+5	1	+5	25	25
89	+6	2	+12	36	72
			−23 + 34		218
			+11		

$$[z] = \tfrac{11}{25} = 0.4, \quad [z^2] = \tfrac{218}{25} = 8.7,$$

$$d = -0.4, \quad d^2 = 0.16,$$

$$c = 83, \quad [z^2] = 8.7.$$

Therefore

$$c - d = a = 83.4, \quad [z^2] - d^2 = [x^2] = 8.56.$$

§ 8

According to equations (7) the expected mean value of the powers of $(a' - a)$ depends upon the values of $\sigma, \sigma_3, \sigma_4$. It is therefore important to determine also their average accuracy.

$$\sigma_r'^r = [x'^r], \quad \sigma_r''^r = [x_r''^r].$$

Therefore

$$[\sigma_r'^r] = [\sigma_r^r].$$

We may estimate here also the average variations of $\sigma_r'^r$ by squaring $\sigma_r'^r - \sigma_r^r$,

$$(\sigma_r'^r - \sigma_r^r)^2 = \left(\frac{x_1^r - \sigma_r^r}{n} + \frac{x_2'^r - \sigma_r^r}{n} + \cdots + \frac{x_n'^r - \sigma_r^r}{n}\right)^2,$$

$$[(\sigma_r'^r - \sigma_r^r)^2] = \frac{[(x^r - \sigma_r^r)^2]}{n} = \frac{[x^{2r}] - 2\sigma_r'^r[x^r] + \sigma_r^{2r}}{n},$$

$$= \frac{\sigma_{2r}^{2r} - \sigma_r^{2r}}{n}; \tag{17}$$

and especially

$$[(\sigma'^2 - \sigma^2)^2] = \frac{\sigma_4^4 - \sigma^4}{n}. \tag{17*}$$

Digitized by Google

If we call $\sigma_r' = \sigma_r + \epsilon_r$ and consider ϵ_r as a small value, we can write

$$\sigma_r'^r - \sigma_r^r = r\epsilon_r\sigma_r^{r-1}$$

and

$$\epsilon_r = \frac{\pm 1}{r\sigma_r^{r-1}}\sqrt{\frac{\sigma_{2r}^{2r} - (\sigma_r^r)^2}{n}}. \tag{18}$$

For σ the value will be

$$\epsilon = \pm\frac{1}{2\sigma}\sqrt{\frac{\sigma_4^4 - \sigma^4}{n}}. \tag{18*}$$

Here σ_r and σ_{2r} are known only approximately and we may estimate more strictly

$$[(x_r'^r - \sigma_r'^r)^2] = [(x_r'^r - \sigma_r^r) - (\sigma_r'^r - \sigma_r^r)]^2 = [(x_r'^r - \sigma_r^r)]^2 - [(\sigma_r'^r - \sigma_r^r)^2],$$

$$\sigma_{2r}'^{2r} - \sigma_r'^{2r} = (\sigma_{2r}^{2r} - \sigma_r^{2r})\frac{n-1}{n}.$$

By substitution in (17)

$$[(\sigma_r'^r - \sigma_r^r)^2] = \frac{\sigma_{2r}'^{2r} - \sigma_r'^{2r}}{n-1}. \tag{19}$$

It is easy to prove in the same way as has been done for $a' - a$, that $\sigma_r'^r - \sigma_r^r$ must be distributed the more nearly according to the exponential law, the greater n is.

§ 9

It was stated on p. 5 that the most convenient way of representing the distribution of the measurements of a variable is by counting the number of measurements that occur within convenient intervals. In many cases this is the only feasible method of taking the measurements.

The question therefore arises as to how far the measurements grouped in such intervals agree with the theoretical series or what corrections have to be made.

If we imagine the series of observations recorded in intervals of $d/2$ and if the number of cases in the interval of $X - d/2$ and X is $f(X - d/4)$, we have, the average obtained under these conditions being designated by $a_{d/2}$,

$$a_{d/2} = \left[\left(X - \frac{d}{4}\right)f\left(X - \frac{d}{4}\right) + \left(X + \frac{d}{4}\right)f\left(X + \frac{d}{4}\right) + \cdots\right].$$

In order to compare this average with the average obtained when

Digitized by Google

intervals d are used, we may write

$$a_d = \left[\left(X-\frac{d}{4}+\frac{d}{4}\right)f\left(X-\frac{d}{4}\right)+\left(X+\frac{d}{4}-\frac{d}{4}\right)f\left(X+\frac{d}{4}\right)+\cdots\right]$$

$$= \left[\left(X-\frac{d}{4}\right)f\left(X-\frac{d}{4}\right)+\left(X+\frac{d}{4}\right)f\left(X+\frac{d}{4}\right)+\cdots\right.$$

$$\left.\cdots+\frac{d}{4}\left\{f\left(X-\frac{d}{4}\right)-f\left(X+\frac{d}{4}\right)\right\}+\cdots\right].$$

The second part of our series is determined by the value of

$$\left[f\left(X-\frac{d}{4}\right)-f\left(X+\frac{d}{4}\right)+\cdots\right]=\left[f\left(X-\frac{d}{4}\right)\right.$$

$$\left.+f\left(X-\frac{d}{4}+\frac{2d}{2}\right)+\cdots\right]-\left[f\left(X-\frac{d}{4}+\frac{d}{2}\right)+f\left(X-\frac{d}{4}+\frac{3d}{2}\right)=\cdots\right].$$

The negative element on the right-hand side of this equation may be expressed by

$$\cdots y_{-2}+y_0+y_{+2}+y_{+4}\cdots.$$

Then provided the interpolation formula (§ 2) is valid

$$y_1 = \cdots + c_{-2}y_{-2} + c_0y_0 + c_2y_2 + c_4y_4 + \cdots,$$

where the values c designate certain constants. By the same formula we find

$$y_3 = \cdots + c_{-2}y_0 + c_0y_2 + c_2y_4 + c_4y_6 + \cdots,$$

and by adding all these values,

$$+y_{-3}+y_{-1}+y_1+y_3+\cdots=(\cdots c_{-2}+c_0+c_2+c_4+\cdots)$$

$$\times(\cdots y_{-2}+y_0+y_2+y_4+\cdots).$$

This presupposes that the frequencies for large positive or negative values of X will be very small. According to our definition of variability, this must always be the case (see p. 5).

We can show in the same way that

$$\cdots+y_{-2}+y_0+y_2+y_4+\cdots=(\cdots c_{-2}+c_0+c_2+c_4+\cdots)$$

$$\times(\cdots y_{-3}+y_{-1}+y_1+y_3+\cdots),$$

and therefore

$$\cdots+y_{-2}+y_0+y_2+y_4+\cdots=\cdots y_{-3}+y_{-1}+y_1+y_3+\cdots.$$

It follows from this that the difference between these two values is

Digitized by Google

zero, and that

$$a_{-d} = a_{d/2}, \tag{20}$$

provided that the interpolation formula adequately represents the frequency function between the limits of any X and $X + d$.

We can calculate in the same way, provided the interpolation formula is applicable,

$$\begin{aligned}\sigma_d^2 &= \left[\left(X - \frac{d}{4} + \frac{d}{4}\right)^2 f\left(X - \frac{d}{4}\right) + \left(X + \frac{d}{4} - \frac{d}{4}\right)^2 f\left(X + \frac{d}{4}\right) + \cdots\right] \\ &= \left[\left(X - \frac{d}{4}\right)^2 f\left(X - \frac{d}{4}\right) + \left(X + \frac{d}{4}\right)^2 f\left(X + \frac{d}{4}\right)\right] \\ &\quad + \frac{d}{2}\left[\left(X - \frac{d}{4}\right) f\left(X - \frac{d}{4}\right) - \left(X + \frac{d}{4}\right) f\left(X + \frac{d}{4}\right) + \cdots\right] \\ &\quad + \frac{d^2}{16}\left[f\left(X - \frac{d}{4}\right) + f\left(X + \frac{d}{4}\right) + \cdots\right].\end{aligned}$$

It has been shown before that the second term of the sum is zero, therefore

$$\sigma_d^2 = \sigma_{d/2}^2 + \frac{d^2}{16}.$$

By continued division we find

$$\sigma_{d/2}^2 = \sigma_{d/4}^2 + \frac{d^2}{64}$$

and

$$\begin{aligned}\sigma_d^2 &= \sigma^2 + \frac{d^2}{16} + \frac{d^2}{64} + \cdots \\ &= \sigma^2 + \frac{d^2}{12}. \qquad (21)\end{aligned}$$

In the same way we find

$$\sigma_{3,d}^3 = \sigma_3^3 \tag{22}$$

and for

$$\begin{aligned}\sigma_{4,d}^4 &= \sigma_{4,d/2}^4 + \tfrac{6}{16}\sigma_{d/2}^2 d^2 + \frac{d^4}{256} \\ &= \sigma_{4,d/2}^4 + \tfrac{3}{8}\sigma^2 d^2 + \frac{3d^4}{256}\end{aligned}$$

and by continued subdivision

$$\sigma_{4,d}^4 = \sigma_4^4 + \tfrac{1}{2}\sigma^2 d^2 + \frac{d^4}{80}$$

$$\sigma_4^4 = \sigma_{4,d}^4 - \tfrac{1}{2}\sigma_d^2 d^2 + \frac{7}{240} d^4. \tag{23}$$

Digitized by Google

For example, for the table on p. 24 we found for interval 1 the value of $\sigma^2 = 8.91$, the corrected value $\sigma_d^2 - \frac{1}{12}d^2$ would be, therefore, $8.91 - 0.08 = 8.83$. The actual results for the calculation of the average and of the mean square variation may be seen from a grouping in larger intervals of the material given on pp. 5 and 6.

1055–1075	1	1045–1065	1
1075–1095	2	1065–1085	2
1095–1115	3	1085–1105	1
1115–1135	6	1105–1125	5
1135–1155	5	1125–1145	6
1155–1175	26	1145–1165	11
1175–1195	48	1165–1185	44
1195–1215	83	1185–1205	52
1215–1235	93	1205–1225	95
1235–1255	121	1225–1245	104
1255–1275	136	1245–1265	130
1275–1295	105	1265–1285	135
1295–1315	106	1285–1305	108
1315–1335	83	1305–1325	89
1335–1355	46	1325–1345	62
1355–1375	21	1345–1365	35
1375–1395	11	1365–1385	14
1395–1415	4	1385–1405	4
1415–1435	2	1405–1425	3
1435–1455	—	1425–1445	1
1455–1475	1	1445–1465	1
1475–1495	1	1465–1485	1
1495–1515	1	1485–1505	1
a	126.50		126.45
σ	±5.64		±5.51

1055–1085	3	1045–1075	1	1035–1065	1
1085–1115	3	1075–1105	3	1065–1095	2
1115–1145	9	1105–1135	8	1095–1125	6
1145–1175	28	1135–1165	14	1125–1155	8
1175–1205	79	1165–1195	65	1155–1185	53
1205–1235	145	1195–1225	136	1185–1215	104
1235–1265	184	1225–1255	171	1215–1245	147
1265–1295	178	1255–1285	198	1245–1275	203
1295–1325	154	1285–1315	149	1275–1305	170
1325–1355	81	1315–1345	110	1305–1335	124
1355–1385	30	1345–1375	40	1335–1365	62
1385–1415	6	1375–1405	13	1365–1395	16
1415–1445	2	1405–1435	4	1395–1425	5
1445–1475	1	1435–1465	1	1425–1455	1
1475–1505	2	1465–1495	1	1455–1485	2
		1495–1525	1	1485–1515	1
a	126.52		126.48		126.42
σ	±5.64		±5.63		±5.63

Digitized by Google

1045–1095	3	1055–1105	4	1015–1065	1	1025–1075	1
1095–1145	12	1105–1155	13	1065–1115	5	1075–1125	8
1145–1195	76	1155–1205	105	1115–1165	20	1125–1175	34
1195–1245	230	1205–1255	266	1165–1215	148	1175–1225	174
1245–1295	308	1255–1305	306	1215–1265	277	1225–1275	307
1295–1345	216	1305–1355	170	1265–1315	284	1275–1325	259
1345–1395	51	1355–1405	34	1315–1365	145	1325–1375	102
1395–1445	6	1405–1455	4	1365–1415	20	1375–1425	16
1445–1495	2	1455–1505	3	1415–1465	3	1425–1475	2
1495–1545	1			1465–1515	2	1475–1525	2
a	126.55		126.51		126.47		126.43
σ	±5.80		±5.70		±5.81		±5.80

or corrected $\sigma = \pm 5.62$

According to (13), the mean square error of our result will be

$$\pm \frac{\sigma}{\sqrt{n-1}} = \pm \frac{5.62}{\sqrt{904}} = \pm 0.19.$$

It appears, therefore, that here, even when our values are combined in groups of 5 cm., the difference between the averages and the average obtained from a finer subdivision is less than the mean square error. Since the accuracy of the results can never be greater than the expected error, there is no need to use a subdivision which is so fine that the mean square error is much larger than the effect of the size of the intervals upon the result. The amount of labor involved in our calculations will depend upon the number of subdivisions of the range, and it is, therefore, desirable to determine the most favorable division of the range before commencing any lengthy computation.

B. Comparison of Limited and Unlimited Series.

§ 10

We have heretofore considered the relation between certain averages of an empirical series and the infinitely long series which it represents. We will now take up the question of how far the empirical series corresponds in detail to the unlimited series. This question may best be approached by assuming the unlimited series to be known, and by investigating how far the limited series corresponds to the unlimited one. If a certain measurement in the unlimited series — which we shall call the type series — occurs with a certain relative frequency, the question arises whether in a limited series of observations we may expect this measurement with the same relative frequency, or whether it is probable that other relative frequencies may be found.

Digitized by Google

To solve this question a number of simple definitions and propositions relating to the theory of probabilities are required.

1. Probability p is called the ratio between the frequency f of an event and the total number of cases n in which the event may occur.

$$p = \frac{f}{n}. \tag{24}$$

If I throw one 10,000 times among 60,000 casts with an ordinary die, the probability p of the throw one is 10,000/60,000. Since the frequency f can never be less than zero nor more than n, it follows that p is always a proper fraction.

It follows from this definition that the frequency is equal to the product of the probability and the number of cases.

$$f = pn. \tag{25}$$

2. If one event (A) has the probability p_1, another (B) the probability p_2, the probability that either the event A or the event B will occur is

$$P = p_1 + p_2. \tag{26}$$

In a number of cases n the event A occurs with the frequency f_1, the event B with the frequency f_2. If the two events are entirely independent, their combined frequency — that is, the frequency with which either one or the other event may be expected — will be $f_1 + f_2$, and their probability, therefore,

$$\frac{f_1 + f_2}{n} = p_1 + p_2.$$

3. If one event has the probability p_1, another entirely independent event the probability p_2, the probability that both will occur jointly

$$P = p_1 p_2. \tag{27}$$

In a number of cases n the two events A and B are to occur. There are f_1 cases in which event A occurs. Among these, event B may occur $f_1 \cdot p_2$ times, so that $f_1 \cdot p_2$ gives us the frequency of all the cases in which both events occur. Their probability is, therefore,

$$\frac{f_1 p_2}{n} = p_1 p_2.$$

Provided a certain event has the probability p, this probability being its relative frequency in an infinitely long series, then the

Digitized by Google

probability that the event does not occur will be $1 - p$. This value may be called q. If we indicate the occurrence of the event by 1, its non-occurrence by 0, then in a series of n observations the following groups of combinations may occur.

The event does not occur at all among n cases

$$000 \cdots 000$$

I. The event occurs once, and does not occur $n - 1$ times.

$$\begin{array}{l} 100 \cdots 000 \\ 010 \cdots 000 \\ 001 \cdots 000 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ 000 \cdots 100 \\ 000 \cdots 010 \\ 000 \cdots 001 \end{array}$$

II. The event occurs twice, and does not occur $n - 2$ times.

$$\begin{array}{l} 110 \cdots 000 \\ 101 \cdots 000 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ 100 \cdots 100 \\ 100 \cdots 010 \\ 100 \cdots 001 \\ 011 \cdots 000 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ 010 \cdots 100 \\ 010 \cdots 010 \\ \text{etc.} \end{array}$$

In the first case, namely when the event does not occur at all, every non-occurrence has the probability q, consequently the probability that the event does not occur n times is, according to (27), q^n.

In Group I. we must consider the various combinations singly. The $n - 1$ cases of non-occurrence have each the probability q; the single case of occurrence has the probability p; consequently every combination has the probability pq^{n-1}.

In the same way we find that in Groups II., III., $\cdots r$, each combination has the probability p^2q^{n-2}, p^3q^{n-3}, $\cdots$ and for Group r

$$P'_r = p^r q^{n-r}. \tag{28}$$

Since, for our particular purpose, we want to know how often among a group of n observations the event occurs r times, the order of occurrence and non-occurrence is irrelevant, so that, according to (26), the total probabilities of r occurrences among n cases is found by adding the probabilities of all the cases in each group. It is obvious that there is only one case when the occurrence does not

Digitized by Google

take place at all. In Group I. the case of occurrence may be in the first, second, $\cdots n$th positions. There are, therefore, n cases contained in this group, so that the probability of one occurrence out of n observations will be npq^{n-1}.

In Group II. it appears, from what has been said before, that taking only those cases where occurrence is in the first place, there will be $n-1$ positions for the second occurrence. Since the first occurrence may be found in all positions from the first to the nth, there are in all $n(n-1)$ cases possible. It is, however, obvious that here those cases have been counted twice in which the first and second occurrences have exchanged positions. For instance, the case where the first occurrence is in first place, the second in second place, is identical with the case where the first occurrence is in second place, the second in first place. The total number is, therefore, only one-half of the total combinations obtained, namely, $n(n-1)/1 \cdot 2$.

We can continue in this manner and find that for Group III. the number of different combinations is $n(n-1)(n-2)/1 \cdot 2 \cdot 3$ and generalized for Group r

$$\frac{n(n-1)(n-2)\cdots(n-r+1)}{1 \cdot 2 \cdots r},$$

and, since the probability of each is, according to (28), $p^r q^{n-r}$, we have the total probability of a combination of r occurrences and $n-1$ non-occurrences.

$$(29) \qquad P_{r,n} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{1 \cdot 2 \cdot 3 \cdots r} p^r q^{n-r}.$$

A better proof of this formula may be given in the following manner. Provided $F_{r,n}$ is the frequency of the combination of r occurrences among n occurrences, then we may determine its frequency among $n+1$ occurrences. The one added case may be in 1st, 2d, 3d, etc., to the $n+1$st position, and for each of these positions $F_{r,n}$ combinations will occur. Thus will originate $(n+1)F_{r,n}$ combinations. In the one added case the event shall not occur. We have, therefore, the combination

$$0(111\cdots000\cdots).$$

In every one of these $n+1$ groups the place of the one additional case of non-occurrence can be taken by the $n-r$ cases of non-occurrence in the original series, so that, including the additional case, there are in each group $n-r+1$ cases in which occurrences and

Digitized by Google

non-occurrences follow in the same order. The total number of combinations is, therefore,

$$F_{r,n+1} = \frac{n+1}{n-r+1} F_{r,n}.$$

The group which contains the case in which the event occurs every time has only one combination; that is, if we take $n = r$,

$$F_{r,r} = 1,$$

$$F_{r,r+1} = \frac{r+1}{1},$$

$$F_{r,r+2} = \frac{(r+2)(r+1)}{1 \cdot 2},$$

and for

$$F_{r,n} = F_{r,r+(n-r)} = \frac{n(n-1)\cdots(r+1)}{1 \cdot 2 \cdots (n-r)}.$$

For $F_{n-r,n}$ we find

$$F_{n-r,n} = F_{n-r,(n-r)+r} = \frac{n(n-1)\cdots(n-r+1)}{1 \cdot 2 \cdots r},$$

By multiplying the value found for $F_{r,n}$ with

$$\frac{r(r-1)\cdots(n-r+1)}{(n-r+1)(n-r+2)\cdots r} = 1$$

we find

$$\frac{n(n-1)\cdots(n-r+1)}{1 \cdot 2 \cdots r} = \frac{n(n-1)\cdots(r+1)}{1 \cdot 2 \cdots (n-r)},$$

therefore

(30) $$F_{n-r,n} = F_{r,n}.$$

Since, according to (28), the probability of each combination of Group r is $p^r q^{n-r}$, we find for the total probability of r occurrences among n observations

(29) $$P_{r,n} = \frac{n(n-1)\cdots(n-r+1)}{1 \cdot 2 \cdots r} p^r q^{n-r}.$$

This probability is equal to the $(r+1)$th term of the binomial expansion $(q+p)^n$; and for this reason the law expressing the probability of r occurrences in a series of n observations of an event which has the probability p is called the binomial law.

§ 11

It is important to investigate a few of the characteristics of the binomial law. First of all we will determine for what value of r the

Digitized by Google

probability reaches a maximum; that is, for what value of r

$$P_{r,n} - P_{r+1,n} > 0$$

and

$$P_{r,n} - P_{r-1,n} > 0$$

$$P_{r,n} - P_{r+1,n} = \frac{n(n-1)\cdots(n-r+1)}{1\cdot 2\cdots r} p^r q^{n-r} - \frac{n(n-1)\cdots(n-r)}{1\cdot 2\cdots(r+1)} p^{r+1} q^{n-r-1}$$

$$= \frac{n(n-1)\cdots(n-r+1)}{1\cdot 2\cdots r} p^r q^{n-r}\left(1 - \frac{n-r}{r+1}\frac{p}{q}\right),$$

$$P_{r,n} - P_{r-1,n} = \frac{n(n-1)\cdots(n-r+1)}{1\cdot 2\cdots r} p^r q^{n-r} - \frac{n(n-1)\cdots(n-r+2)}{1\cdot 2\cdots(r-1)} p^{r-1} q^{n-r+1}$$

$$= \frac{n(n-1)\cdots(n-r+1)}{1\cdot 2\cdots r} p^r q^{n-r}\left(1 - \frac{r}{n-r+1}\frac{q}{p}\right),$$

therefore

$$P_{r,n} - P_{r+1,n} > 0,$$

when

$$1 - \frac{n-r}{r+1}\frac{p}{q} > 0$$

or

$$r > np - q;$$

and

$$P_{r,n} - P_{r-1,n} > 0,$$

when

$$1 - \frac{r}{n-r+1}\frac{q}{p} > 0$$

or

$$r < np - q + 1;$$

that is to say, r is the integer between $np - q$ and $np - q + 1$. If np is an integer, this value of r

$$r_{max} = np.$$

If np is not an integer, it is the nearest integer under or over np. Thus we have found that the greatest probability belongs to that frequency of the event which corresponds most accurately to its probability.

§ 12

In order to gain a better insight into the distribution of those frequencies which differ from the typical frequency, we will compare

Digitized by Google

those values which are equidistant from the maximum frequency, which, for convenience sake, may be considered an integer

$$r_1 = np + s,$$

$$r_2 = np - s,$$

$$P_{r_1,n} = \frac{n(n-1)\cdots(n-np-s+1)}{1\cdot 2\cdots(np+s)} p^{np+s} q^{n-np-s}$$

$$= \frac{n(n-1)\cdots(nq-s+1)}{1\cdot 2\cdots(np+s)} p^{np+s} q^{nq-s},$$

$$P_{r_2,n} = \frac{n(n-1)\cdots(nq+s+1)}{1\cdot 2\cdots(np-s)} p^{np-s} q^{nq+s},$$

$$\frac{P_{r_1,n}}{P_{r_2,n}} = \frac{(nq+s)(nq+s-1)\cdots(nq-s+1)p^{2s}}{(np+s)(np+s-1)\cdots(np-s+1)q^{2s}}$$

$$= \frac{\left(1+\frac{s}{nq}\right)\left(1+\frac{s-1}{nq}\right)\cdots\left(1-\frac{s-1}{nq}\right)}{\left(1+\frac{s}{np}\right)\left(1+\frac{s-1}{np}\right)\cdots\left(1-\frac{s-1}{np}\right)}.$$

If we consider s as so small in comparison to np and nq that the fractions may be disregarded, the value is nearly equal to one; that is, the distribution will be nearly symmetrical. If we assume s as so small in comparison to np and nq, that the higher powers of s/np and s/nq may be disregarded, we have

$$\frac{P_{r_1,n}}{P_{r_2,n}} = \frac{1+\frac{s+(s-1)+\cdots[-(s-1)]}{nq}}{1+\frac{s+(s-1)r\cdots[-(s-1)]}{np}}$$

$$= \frac{1+\frac{s}{nq}}{1+\frac{s}{np}}$$

$$= 1+\frac{s}{nq}-\frac{s}{np}$$

$$= 1+\frac{s(p-q)}{npq}.$$

It appears from this equation that the degree of asymmetry will be the same for large values of np and nq as long as the proportion

Digitized by Google

of s and npq remains the same. It also appears that the more nearly p and q are equal the greater will be the symmetry of the whole series.

We may also determine the same proportion when we consider the next higher powers of s/nq and s/np. We have then

$$\frac{P_{r_1,n}}{P_{r_2,n}} = \frac{1 + \frac{s}{nq} + \frac{s[(s-1)+(s-2)+\cdots(-s+1)]}{n^2q^2} + (s-1)[s+(s-2)+\cdots(-s+1)] + \cdots}{1 + \frac{s}{np} + \frac{s[(s-1)+(s-2)+\cdots(-s+1)]}{n^2p^2} + (s-1)[(s-2)+\cdots(-s+1)] + \cdots}.$$

It can easily be shown that the third terms in numerator and denominator

$$\Sigma = -\frac{(s-1)s(2s-1)}{1.2.3}.$$

Thus we find

$$\frac{P_{r_1,n}}{P_{r_2,n}} = \frac{1 + \frac{s}{nq} - \frac{(s-1)s(2s-1)}{1.2.3nq^2}}{1 + \frac{s}{np} - \frac{(s-1)s(2s-1)}{1.2.3np^2}}$$

$$= 1 + \frac{(p-q)}{npq}\left(s - \frac{(s-1)s(2s-1)}{1.2.3npq}\right),$$

or, if

$$\frac{s^2}{(npq)^2} \quad \text{and} \quad \frac{s}{(npq)^2}$$

are disregarded

$$= 1 + \frac{p-q}{npq}\left(s - \frac{s^3}{3npq}\right). \tag{31}$$

§ 13

We will next estimate the probability of the event occurring less than s times among n times.

$$P_s = \frac{n(n-1)\cdots(n-s+1)}{1\cdot 2\cdots s} p^s q^{n-s},$$

$$P_s < n^s p^s q^{n-s},$$

$$< \left(\frac{np}{q}\right)^s q^n.$$

It follows from this that the sum of all the probabilities of the event

Digitized by Google

occurring $0, 1, 2 \cdots s$ times

$$\sum_0^s (P_s) < \frac{\left(\frac{np}{q}\right)^{s+1} - 1}{\frac{np}{q} - 1} q^n.$$

It is easy to show that the numerator decreases with increasing n.

$$\left(\frac{np}{q}\right)^{s+1} q^n : \left(\frac{\overline{n+1}p}{q}\right)^{s+1} q^{n+1} = 1 : q\left(1 + \frac{1}{n}\right)^{s+1}.$$

Since q is a fraction, the last product will be a fraction as soon as n is taken sufficiently large. It follows that from a certain value of n up, the product $(np/q)^{s+1}q^n$ becomes smaller and smaller: therefore the value $\Sigma_0^s(P_s)$ will also become smaller and smaller with increasing values of n.

§ 14

We will finally compare the frequency at the point $np + s_1$ with that at the point $np + s_2$.

$$P_{np+s_1} = \frac{n(n-1)\cdots(nq - s_1 + 1)}{1 \cdot 2 \cdots (np + s_1)} p^{np+s_1} q^{nq-s_1},$$

$$P_{np+s_2} = \frac{n(n-1)\cdots(nq - s_2 + 1)}{1 \cdot 2 \cdots (np + s_2)} p^{np+s_2} q^{nq-s_2},$$

$$\frac{P_{np+s_2}}{P_{np+s_1}} = \frac{(nq - s_1)(nq - s_1 - 1)\cdots(nq - s_2 + 1)}{(np + s_1 + 1)(np + s_1 + 2)\cdots(np + s_2)} \frac{p^{s_2 - s_1}}{q^{s_2 - s_1}}$$

$$= \frac{\left(1 - \frac{s_1}{nq}\right)\left(1 - \frac{s_1 + 1}{nq}\right)\cdots\left(1 - \frac{s_2 - 1}{nq}\right)}{\left(1 + \frac{s_1 + 1}{np}\right)\left(1 + \frac{s_1 + 2}{np}\right)\cdots\left(1 + \frac{s_2}{np}\right)}.$$

If we assume that $(s_2^2 - s_1^2)/np$ and $(s_2^2 - s_1^2)/nq$ are so small that their higher powers may be disregarded, we have

$$\frac{P_{np+s_2}}{P_{np+s_1}} = 1 - \frac{(s_2 - s_1)(s_1 + s_2 - 1)}{2nq} - \frac{(s_2 - s_1)(s_1 + s_2 + 1)}{2np}$$

$$= 1 - \frac{s_2^2 - s_1^2}{2npq} - \frac{(s_2 - s_1)(p - q)}{2npq}.$$

We will call

$$s_2 - s_1 = d,$$

Digitized by Google

and assume

$$\varepsilon_1 = rd.$$

$$\frac{P_{np+(r+1)d}}{P_{np+rd}} = 1 - \frac{d^2(2r+1)}{2npq} - \frac{d(p-q)}{2npq}.$$

We will add all the probabilities for the points $np + rd + 1$, $np + rd + 2, \cdots np + (r+1)d$.

$$\frac{P_{np+rd+1} + P_{np+rd+2} + \cdots P_{np+r+1d}}{P_{np+rd}} = d - \frac{(1+2^2+3^2+\cdots d^2)(2r+1)}{2npq} - \frac{(1+2+\cdots d)(p-q)}{2npq}$$

$$= d - \frac{d(d+1)(2d+1)(2r+1)}{1\cdot 2\cdot 3\cdot 2npq} - \frac{d(d+1)(p-q)}{1\cdot 2\cdot 2npq}$$

$$= d\left[1 - \frac{d^2(2r+1)}{6npq} - \frac{d(r+p)}{2npq} + \frac{2r+1+3(p-q)}{12npq}\right].$$

The ratio between the second and third terms of this expression

$$\frac{d^2(2r+1)}{6npq} : \frac{d(r+p)}{2npq} = d : \frac{3(r+p)}{2r+1}$$

and the value of this ratio lies between

$$d : 3p \text{ (for } r = 0) \quad \text{and} \quad d : \tfrac{3}{2} \text{ (for } r = \infty).$$

In the same way the ratio between the second and last terms can be shown to be between

$$d^2 : \tfrac{1}{2} + \tfrac{3}{2}(p-q) \quad \text{and} \quad d^2 : \tfrac{1}{2}.$$

Therefore if d is taken sufficiently large, which implies also an n sufficiently large, the third and fourth terms may be disregarded, and, by designating the sum of the probabilities between rd and $(r+1)d$ by P_r, we have—designating the sum of the values P_r by ΣP_r—

$$\sum_{np+rd}^{np+(r+1)d} P_r = P_{np+rd} d\left[1 - \frac{d^2(2r+1)}{6npq}\right]. \tag{32}$$

Since, according to § 13, all the terms beyond sufficiently large limits of rd can be made as small as we desire, we can say

$$\Sigma P_r = 1 = \Sigma P_{np+rd} d\left[1 - \frac{d^2(2r+1)}{6npq}\right].$$

If we assume

$$d = c\sqrt{npq},$$

Digitized by Google

we find that

$$\Sigma P_{np+rc\sqrt{npq}} c\sqrt{npq}\left[1-\frac{c^2(2r+1)}{6}\right]=1.$$

In the same way we find for a second series

$$\Sigma P'_{n'p'+rc\sqrt{n'p'q'}} c\sqrt{n'p'q'}\left[1-\frac{c^2(2r+1)}{6}\right]=1,$$

and

$$\Sigma P_{np+rc\sqrt{npq}} c\sqrt{npq}=\Sigma P'_{n'p'+rc\sqrt{n'p'q'}} c\sqrt{n'p'q'}.$$

Since this equation must hold for any value of r and c each

$$P_{np+rc\sqrt{npq}} c\sqrt{npq}=P'_{n'p'+rc\sqrt{n'p'q'}} c\sqrt{n'p'q'},$$

or

$$P_{np+rc\sqrt{npq}} : P'_{n'p'+rc\sqrt{n'p'q'}} = \sqrt{n'p'q'} : \sqrt{npq}. \tag{33}$$

In other words, for sufficiently large values of np and nq the probabilities for points removed from the most probable point by equal multiples of $\sqrt{npq}$ are inversely proportional to this value, and therefore equal for series in which $\sqrt{npq}$ has the same value.

According to (32) we have, by substituting for d, $c\sqrt{npq}$,

$$\sum_{np+rd}^{np+\overline{r+1}d} P_r = P_{np+rc\sqrt{npq}}\, c\sqrt{npq}\left[1-\frac{c^2(2r+1)}{6}\right].$$

According to (33) this value is constant.

In other words, for sufficiently large values of np and nq the total probability between two limits is always the same when the limits are the same multiples of $\sqrt{npq}$. This value is, therefore, the standard by which the probabilities of deviations from the normal frequency are measured, and is called the standard deviation, and may be designated by σ.

We may now summarize the results so far obtained. We have found that when a limited series of n observations is taken, which is a representative of an infinitely long series in which a certain event occurs with the probability p, then the most probable frequency of the event will be np. Deviations from this frequency are found with frequencies corresponding to the binomial law. By a series of approximations it has been found that, if np and nq are large values, the distribution of frequencies will be very nearly symmetrical around the value np, and the frequencies of points far removed from the value np will be negligible. In this case frequencies of deviations from np will be, in various series, inversely proportional to $\sqrt{npq}$, when

Digitized by Google

the deviations are the same multiples of $\sqrt{npq}$; and the total probability of finding any deviation between two limits will be the same as long as the limits remain the same multiples of $\sqrt{npq}$.

If, therefore, a table is computed for the binomial terms, the probability of any deviation, and also the probability of finding any deviation inside of certain limits, can be determined. For large values of $\sqrt{npq}$ a single table will be sufficient.

For our particular purpose it is desired to know how probable it will be that the frequency of an event will lie within certain limits deviating from the desired value np. We may wish to determine how often, in a number of series of n observations, we may expect a frequency that lies between $np - x$ and $np + x$. If, for instance, it can be determined that among 1,000 such series the frequency can be expected only five times to fall outside of these limits, we may conclude that it is reasonably certain that the value p, which we can not determine by direct observation, lies between these limits. For large values of np and nq it has been calculated that the probability of finding a frequency between $np - x$ and $np + x$ is

$$0.632 \text{ for } x = \sqrt{npq},$$

$$0.954 \text{ for } x = 2\sqrt{npq},$$

$$0.997 \text{ for } x = 3\sqrt{npq}.$$

It appears, therefore, that in the last case we may be almost certain that the true value will not differ from the empirical value obtained from n observations by more than $3\sqrt{npq}$. Since the probability of this difference is determined by the measure $\sqrt{npq}$, we may call this value also the standard error and designate it, as has been done in the preceding pages by ϵ.

The following table of probabilities of deviations[1] exhibits the degree of accuracy of the approximation discussed in §§ 11–14 when the value of n is not large:

$p = \frac{1}{2}$	$n = 36$	$\sqrt{npq} = 3$	$p = \frac{1}{2}$	$n = 100$	$\sqrt{npq} = 5$
	Probabilities.			Probabilities.	
No. of Occurrences.	Approximation.	Binomial Formula.	No. of Occurrences.	Approximation.	Binomial Formula.
18	0.132	0.132	50	0.080	0.080
17–19	0.383	0.382	49–51	0 236	0.236
15–21	0.757	0.757	45–55	0.729	0.729
12–24	0.970	0.971	40–60	0.964	0.965
9–27	0.998	0.999	35–65	0.998	0.998
6–30	0.99997	0.99999	30–70	0.99996	0.99997

When p and q differ, the approximation is not quite so good.

[1] H. Westergaard, 'Die Grundzüge der Theorie der Statistik,' pp. 68, 69.

Digitized by Google

$p=\frac{1}{10}$	$n=100$	$\sqrt{npq}=3$	$p=\frac{1}{10}$	$n=1000$	$\sqrt{npq}=9.5$
	Probabilities.			Probabilities.	
No. of Occurrences.	Approximation.	Binomial Formula.	No. of Occurrences.	Approximation.	Binomial Formula.
10	0.132	0.132	100	0.042	0.042
6–14	0.866	0.870	90–110	0.732	0.732
1–19	0.998	0.998	80–120	0.970	0.969

According to what has been said before, the approximation gives a symmetrical formula, while the binomial formula is the more asymmetrical the smaller the value of n. If we calculate the probabilities of positive and negative groups separately, according to both formulas, these differences appear clearly.[1]

$p=\frac{1}{10}$	$n=100$	$\sqrt{npq}=3$	$p=\frac{1}{10}$	$n=1000$	$\sqrt{npq}=9.5$
	Probabilities.			Probabilities.	
No. of Occurrences.	Approximation.	Binomial Formula.	No. of Occurrences.	Approximation.	Binomial Formula.
10	0.132	0.132	100	0.042	0.042
below 10	0.434	0.451	below 100	0.479	0.485
above 10		0.417	above 100		0.473
6–9	0.367	0.394	90–99	0.345	0.351
11–14		0.344	101–110		0.339
1–5	0.066	0.057	80–89	0.119	0.120
15–19		0.071	111–120		0.117
			60–79	0.015	0.014
			121–140		0.017

This asymmetry may be taken into account by using (31). If we assume that the degree of asymmetry is small, and call $P_{x,n}$ the probability of the point x for a series of n observations, P_x its probability for a series of very many observations, we may write

$$P_{x,n} = P_x(1+d), \qquad P_{-x,n} = P_x(1-d). \qquad \frac{P_{x,n}}{P_{-x,n}} = (1+2d).$$

According to (31)

$$\frac{P_{x,n}}{P_{-x,n}} = 1 + \frac{p-q}{npq}\left(x - \frac{x^3}{3npq}\right),$$

therefore

$$d = \frac{p-q}{2npq}\left(x - \frac{x^3}{3npq}\right).$$

§ 15

We will now apply our considerations relating to the averages of powers (§§ 5–7) to the binomial law. We have in this case

[1] H. Westergaard, 'Die Grundzüge der Theorie der Statistik,' p. 69.

Digitized by Google

$$[X^r] = \Sigma X^r \frac{n(n-1)\cdots(n-X+1)}{1\cdot 2\cdots X} p^X q^{n-X}$$

$$= np\Sigma X^{r-1} \frac{n(n-1)(n-2)\cdots(n-1-X+2)}{1\cdot 2\cdots(X-1)} p^{X-1} q^{(n-1)-(X-1)}.$$

If we call

$$X - 1 = v,$$

$$[X^r] = np\Sigma(1+v)^{r-1} \frac{(n-1)(n-2)\cdots(n-1-v+1)}{1\cdot 2\cdots v} p^v q^{n-1-v}$$

$$= np\left\{\begin{array}{l} 1 + (r-1)\Sigma v \dfrac{(n-1)(n-2)\cdots(n-1-v+1)}{1\cdot 2\cdots v} p^v q^{n-1-v} \\ + \dfrac{(r-1)(r-2)}{1\cdot 2}\Sigma v^2 \dfrac{(n-1)(n-2)\cdots(n-1-v+1)}{1\cdot 2\cdots v} p^v q^{n-1-v} \\ + \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \\ + \Sigma v^{r-1} \dfrac{(n-1)(n-2)\cdots(n-1-v+1)}{1\cdot 2\cdots v} p^v q^{n-1-v} \end{array}\right\}$$

In this manner the successive values of $[X^r]$ may be calculated. If we designate the number of cases from which the average is derived by an inferior mark, we have

$$[X^r]_n = np\left\{1 + (r-1)[X]_{n-1} + \frac{(r-1)(r-2)}{1\cdot 2}[X^2]_{n-1} + \cdots + [X^{r-1}]_{n-1}\right\}.$$

Thus the following values are found

$$(34)\quad \left\{\begin{array}{l} [X]_n = np, \\ [X^2]_n = np + n(n-1)p^2, \\ [X^3]_n = np + 3n(n-1)p^2 + n(n-1)(n-2)p^3, \\ [X^4]_n = np + 7n(n-1)p^2 + 6n(n-1)(n-2)p^3 \\ \qquad\qquad + n(n-1)(n-2)(n-3)p^4, \\ [X^5]_n = np + 15n(n-1)p^2 + 25n(n-1)(n-2)p^3 \\ \qquad + 10n(n-1)(n-2)(n-3)p^4 \\ \qquad\qquad + n(n-1)(n-2)(n-3)(n-4)p^5, \\ [X^6]_n = np + 31n(n-1)p^2 + 90n(n-1)(n-2)p^3 \\ \qquad + 65n(n-1)(n-2)(n-3)p^4 \\ \qquad\qquad + 15n(n-1)(n-2)(n-3)(n-4)p^5 \\ \qquad\qquad\qquad + \{n(n-1)\cdots(n-5)\}p^6. \end{array}\right.$$

Digitized by Google

From these values the average of the powers around the average value of $[X]$ can easily be found by applying (16) or the corresponding formula, developed from

$$[x^r] = [(X-a)^r],$$

and we find

$$(35)\quad \begin{cases} [x] = 0, \\ [x^2] = npq, \\ [x^3] = npq(q-p), \\ [x^4] = 3\overline{npq}^2 + npq(1-6pq), \\ [x^5] = \{10\overline{npq}^2 + npq(1-12pq)\}(q-p), \\ [x^6] = 5.3\overline{npq}^3 + 5\overline{npq}^2(5-26pq) + npq(1-30pq+120p^2q^2). \end{cases}$$

If we determine these values in relation to the total number of cases n, we obtain

$$\left[\frac{x}{n}\right] = 0, \qquad \left[\left(\frac{x}{n}\right)^2\right] = \frac{pq}{n}, \qquad \left[\left(\frac{x}{n}\right)^3\right] = \frac{1}{n}\cdot\frac{pq}{n}\cdot(q-p),$$

$$\left[\left(\frac{x}{n}\right)^4\right] = 3\left(\frac{pq}{n}\right)^2 + \frac{1}{n^2}\frac{pq}{n}(1-6pq)$$

$$\left[\left(\frac{x}{n}\right)^5\right] = \left\{\frac{1}{n}10\left(\frac{pq}{n}\right)^2 + \frac{1}{n^3}\frac{pq}{n}(1-12pq)\right\}(q-p),$$

$$\left[\left(\frac{x}{n}\right)^6\right] = 5.3\left(\frac{pq}{n}\right)^3 + \frac{1}{n^2}\left(\frac{pq}{n}\right)^2(5-26pq) + \frac{1}{n^4}\cdot\frac{pq}{n}(1-30pq+120p^2q^2).$$

If n is assumed sufficiently large we find thus the approximations

$$(36)\quad \begin{cases} \left[\frac{x}{n}\right] = 0, \\ \left[\left(\frac{x}{n}\right)^2\right] = \left(\frac{pq}{n}\right), \\ \left[\left(\frac{x}{n}\right)^3\right] = 0, \\ \left[\left(\frac{x}{n}\right)^4\right] = 3\left(\frac{pq}{n}\right)^2, \\ \left[\left(\frac{x}{n}\right)^5\right] = 0, \\ \left[\left(\frac{x}{n}\right)^6\right] = 5.3\left(\frac{pq}{n}\right)^3. \end{cases}$$

Digitized by Google

The considerations relating to special averages made in § 5 can easily be modified so as to express the sum of a number of special values representing a function. If a is the general average of a function and S_n the sum of n observed values

$$[(S_n - na)^{2p}]^n = [(x_1' + x_2' + \cdots x_n')^{2p}].$$

The expansion of this term (see pp. 18, *et seq.*) leads to the formula

$$P(x) = \frac{1}{\sigma\sqrt{n}\sqrt{2\pi}} e^{-\frac{(S_n - na)^2}{2n\sigma^2}}. \tag{37}$$

Now we will consider as our function the theoretical distribution of chance occurrences of the probability p in a series containing n observations. We have found that these are distributed according to the binomial law, that their average is np (34), and their mean square deviation npq (35). If we consider the sum of m such series, each consisting of n observations, we find, according to (37),

$$P(S_n - mnp) = \frac{1}{\sqrt{mnpq}\sqrt{2\pi}} e^{-\frac{(S_n - mnp)^2}{2mnpq}}.$$

The deviations $S_n - mnp$ may, however, be also considered as the theoretical distribution of chance occurrences of the probability p in a series containing nm observations, and as such follow the binomial law. From this it follows that the binomial law for a series containing a sufficiently large number of cases may be adequately expressed by the exponential formula, in which we substitute for σ the value $\sqrt{npq}$, the mean square of the deviations of the terms of the binomial law from their average.

It is easy to show that the various approximations which we developed in §§ 11–14 hold good for the exponential formula. The probabilities are distributed symmetrically, because their values depend upon x^2 alone. Between two given limits, s_1 and s_2, we find the number of cases

$$\int_{s_1}^{s_2} \frac{1}{\sqrt{npq}\sqrt{2\pi}} e^{-\frac{x^2}{2npq}} dx = \frac{1}{\sqrt{2\pi}} \int_{\frac{x_1}{\sqrt{npq}}}^{\frac{x_2}{\sqrt{npq}}} e^{-\frac{z^2}{2}} dz.$$

The value of this integral will be the same whenever x_1 and x_2 are the same multiple of $\sqrt{npq}$.

The exponential law gives us a means of calculating the frequencies inside of certain multiples of σ. A full and convenient table of the values of these frequencies has been published by W. F.

Digitized by Google

Sheppard.[1] For our present purposes the following data are sufficient. For values of x/σ, the probability of occurrences of a deviation between $-\infty$ and x/σ will be as follows:

$\frac{x}{\sigma}$	Probability.	$\frac{x}{\sigma}$	Probability.
.00	.5000	1.00	.8413
.05	.5199	1.10	.8643
.10	.5398	1.20	.8849
.15	.5596	1.30	.9032
.20	.5793	1.40	.9192
.25	.5987	1.50	.9332
.30	.6179	1.60	.9452
.35	.6368	1.70	.9554
.40	.6554	1.80	.9641
.45	.6736	1.90	.9713
.50	.6915	2.00	.9772
.55	.7088	2.20	.9861
.60	.7257	2.40	.9918
.65	.7422	2.60	.9953
.70	.7580	2.80	.9974
.75	.7734	3.00	.99865
.80	.7881	3.20	.99931
.85	.8023	3.40	.99966
.90	.8159	3.60	.99984
.95	.8289	3.80	.99993
		4.00	.99997

The probabilities for negative values of x/σ may be found by subtracting the values given in the table from 1.

The asymmetry of distribution may here be introduced by the use of (12). When we consider that, according to (35), $\sigma_3^3 = npq(q-p)$,

$$y = \frac{1}{\sqrt{npq}\,\sqrt{2\pi}} e^{-\frac{x^2}{2npq}} \left\{1 + \frac{p-q}{2npq}\left(x - \frac{x^3}{3npq}\right)\right\}.$$

This agrees with the asymmetry found at the end of § 14 (p. 45).

[1] 'New Tables of the Probability Integral' (*Biometrika*, Vol. II., pp. 174, *et seq.*).

Digitized by Google

III. DISTRIBUTION OF VARIABLES AND OF CHANCE VARIATIONS

§ 16

We have shown that the averages of observed series of variables have certain definite relations to the averages of the unlimited series, whatever the law of distribution may be. We have also seen that the law of distribution must always be of such character that the frequency is zero for values far removed from the bulk of the observations. The empirical investigation of great numbers of variables representing many different kinds of phenomena has demonstrated that in a great many cases their law of distribution conforms quite nearly with the exponential law discussed before; that the distribution is, therefore, the same as though there were chance deviations from a certain expected result, the number of contributory elements being very great.

We can imagine such a result to be brought about by the action of a great many contributory causes which affect the values of our observations. Provided that there are n such causes, each increasing the value of our measurement by one small unit, and that the probability of each cause coming into action be p, the conditions will be the same as those discussed in the preceding chapter; and the result must be a distribution corresponding to the binomial law, if n is small; to the exponential law, if n is large. This assumption is merely a simplification of the existence of many small causes which act according to various laws and with varying probabilities; and the agreement between the two will be the better, the greater the number of causes.

A more general proof of the applicability of the exponential law will be given here. We assume that there are a considerable number, n, of causes, each of which has a slight influence upon the observed phenomenon. The functions expressing the effect of each cause upon the phenomenon are unknown. Any deviation from the average will then be

$$x = (d' + y') + (d'' + y'') + \cdots (d^N + y^N)$$

where the values $A + d$ are the averages for each individual con-

Digitized by Google

tributory cause, and y the deviations from the average for each of these contributory causes. If these are entirely independent,

$$[x^2] = ([d'^2] + [y'^2]) + ([d''^2] + [y''^2]) + \cdots ([d^{N^2}] + [y^{N^2}]) = \sigma^2.$$

If the values d^2 and y^2 are all of the same order of smallness, their order — since σ^2 is a finite value — may be indicated by writing

$$d = \frac{\eta}{\sqrt{n}}$$

$$y = \frac{\xi}{\sqrt{n}}.$$

We will call

$$[(d' + y')^2] = \frac{[(\eta' + \xi')^2]}{n} = \frac{s_2'^2}{n}$$

$$[(d'' + y'')^2] = \frac{[(\eta'' + \xi'')^2]}{n} = \frac{s_2''^2}{n}$$

. . . . . .

Since the individual contributory causes are distributed according to a great variety of laws, it seems justifiable to assume that any group of $n - 2p$ functions expressing these laws is on the average the same as any other group of $n - 2p$ functions, provided $n - 2p$ is a large number

$$[(d' + y')^2] + [(d'' + y'')^2] + \cdots [(d^{N-2P} + y^{N-2P})^2] = (n - 2p)\frac{[s_2^2]}{n},$$

and also in general

$$[(d' + y')^r] + [(d'' + y'')^r] + \cdots [(d^{N-2P} + y^{N-2P})^r] = (n - 2p)\frac{[s_r^r]}{n^{r/2}}.$$

If n is large and $r > 2$ this value will approach zero, since all the values of s_r may be assumed to be of the same order (see p. 19).

In its present form our problem assumes the same form as that of the distribution of special averages of one function which has been discussed in §§ 5–9 (pp. 14, *et seq.*). The value σ_r^r is represented in this case by the average of the corresponding values of the component functions. In this case our approximation which results in the exponential formula will hold good as long as $n - 2p$ is a large number, and as long as p is small as compared with n. For cases in which these conditions are not fulfilled asymmetrical distributions may be expected.

Digitized by Google

Formula (12) (p. 21), which determines the skewness of the curve, shows that $[x^3]$ depends upon s_3, and $[x^4]$ upon s_3 and s_4, and, when higher degrees of asymmetry are considered, the values s_5, s_6, etc., will also exert their influence. Since these values depend upon the character of the component functions, the constants of asymmetry will change according to these functions.

It has often been assumed that skew curves will follow the approximation of the binomial law, but, owing to the independence of these constants such results can not be expected. If asymmetries of the order $1/\sqrt{n}$ are considered we may call

$$[s_2^2] = n'pq,$$

$$\frac{[s_3^3]}{[s_2^2]\sqrt{n}} = p - q.$$

Then, according to (12),

$$[x^3] = n'pq(q - p),$$

which, in (35), has been shown to be the mean cube deviation for the binomial law. In this case a certain form of the binomial law can be found, which will correspond to the skewness of the function. The relation between s_2, s_3 and n and p is, however, entirely artificial. If the skewness of the order $1/n$ is taken into consideration, this relation no longer exists. Skew distributions do not correspond, therefore, ordinarily to the binomial law.

It can also be shown that whenever $\sqrt{[x^2]}$ is large as compared with the average A, skew distributions may be expected. According to our definition of variability, the values of the variates must be limited and, since this must be true also of the variable when subject to all the contributory causes which we have assumed to be of the same order, it follows that the average limit of the contributory functions must be less than A/n. Therefore,

$$[x^2] < \frac{A^2}{n},$$

or

$$\sqrt{[x^2]} < \frac{A}{\sqrt{n}}.$$

Therefore, if $\sqrt{[x^2]}$ is great as compared with A, n cannot be large and we must expect skew distributions.

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google